Improving natural language processing with human data

Eye tracking and other data sources reflecting cognitive text processing

Barrett, Maria

Publication date:
2018

Document version
Publisher's PDF, also known as Version of record

IMPROVING NATURAL LANGUAGE PROCESSING WITH HUMAN DATA

MARIA BARRETT

UNIVERSITY OF COPENHAGEN

Eye tracking and other data sources reflecting cognitive text processing

Department of Nordic Studies and Linguistics
Faculty of Humanities
University of Copenhagen

July 2018

Maria Barrett: *Improving natural language processing with human data: Eye tracking and other data sources reflecting cognitive text processing*

Supervisor
Anders Søgaard, Prof., Dr. Phil., PhD
Department of Computer Science, Faculty of Science, University of Copenhagen

Affiliation
Centre for Language Technology, Department of Nordic Studies and Linguistics, Faculty of Humanities, University of Copenhagen

Submitted
Juli 30, 2018

ABSTRACT

When humans perform everyday tasks like reading, speaking, and writing, they cognitively also complete many of the tasks that natural language processing strives for computers to replicate. The traces of human cognitive processing can be collected in various data sources such as eye tracking during reading, keystroke logs from typing and acoustic cues, where milliseconds matter.

We successfully improve supervised cross-domain part-of-speech tagging and parsing using eye-tracking data. We also present first evidence that we can impove weakly supervised part-of-speech induction for English and French using eye-tracking features from reading. We furthermore demonstrate transfer across related languages by using English eye-tracking recordings to improve French part-of-speech induction. The English experiment utilizes our Universal Dependency annotation layer on top of the – at the time – largest available eye-tracking corpus. For part-of-speech induction, eye-tracking data can further be supported by other sources of human features reflecting the cognitive processing of text, such as acoustic features and features from keystroke logs.

We present a novel way of regularizing the attention of a recurrent neural network with human attention derived from gaze features. We utilize the inductive bias from human attention to consistently improve a range of sequence classification tasks, such as detection of abusive language, grammatical error detection, and sentiment classification.

Technology for recording keystroke logs and prosody features is already common. And the recent advancements of low-cost eye tracking technology promise eye-tracking data to be available in larger quantities, also for low-resource languages. Real-world eye-tracking data poses new challenges compared to laboratory data. One study in this thesis presents first evidence that despite the noise and idiosyncrasies, real-world reading data recorded with a consumer-grade eye tracker can be modeled in machine learning models.

In this thesis, we show that there is an unused potential for utilizing eye tracking and other data sources reflecting human cognitive processing of text for natural language processing.

RESUMÉ

Hverdagsaktiviteter såsom at læse, skrive og tale afspejler menneskers kognitive processering af tekst. Mennesker løser ubevidst opgaver, som natursprogsprocessing prøver at få computere til at efterligne. Spor efter den menneskelige processering af tekst kan blive opsamlet igennem øjenbevægelser fra læsning, tastaturtryk fra skrivning og lydoptagelser af tale. Millisekunders forskel afslører væsentlige forskelle.

Vi forbedrer superviseret opmærkning af ordklasse og syntaks på tværs af tekstdomæner ved hjælp af øjenbevægelsesdata. Vi er også de første til at forbedre ordklasseopmærkning uden brug af opmærket træningsdata for engelsk og fransk. Vi viser endda at signalet bliver overført mellem beslægtede sprog, således at vi kan bruge engelsk øjenbevægelsesdata til at forbedre fransk ordklasseopmærkning. Til de engelske ordklaseeksperimenter bruger vi vores egen *Universal Dependencies*-annotering af det – på den tid – største øjenbevægelseskorpus. Øjenbevægelsesdata kan yderligere blive hjulpet ved at blive kombineret med andre datatyper, der også afspejler kognitiv processering af tekst. Vi kombinerer med akustisk data og tastaturtryk til ordklasseopmærkning.

Vi præsenterer også en ny måde at vægte hvert ord ved hjælp af øjenbevægelser i et dybt neuralt netværk. Dette bruger vi til konsistent at forbedre en række tekstklassificeringsopgaver. Vi opfanger diskriminerende sprog, grammatiske fejl samt holdninger/følelser udtrykt i teksten.

Teknologi til at optage tastaturtryk og tale er allerede udbredt. Og de nyeste opfindelser har også gjort *eyetrackere* til at betale, således at vi må forvente at kunne få adgang til større mængder øjenbevægelsesdata, selv for sprog hvor der normalt ikke findes opmærkede tekstressourcer. Øjenbevægelsesdata fra den virkelige verden introducerer nye udfordringer i forhold til data optaget i laboratorier. Et studie i denne afhandling viser, at vi kan modellere øjenbevægelsesdata fra den virkelige verden optaget med en billigere eyetracker i maskinlæringsmodeller på trods af datastøj og idiosynkrasier.

Denne afhandling viser, at der er et uudnyttet potentiale for natursprogsprocessering i forhold til at bruge øjenbevægelsesdata og andre datakilder, der afspejler kognitiv processing af tekst.

PUBLICATIONS

THESIS STRUCTURE

This is an article-based PhD thesis. Part i will provide an overview of the field of natural language processing (NLP) enriched with human data as well as a brief introduction to what type of information we can expect from eye-tracking data. Each of the eight chapters presented in Part ii, Part iii, and Part iv represents a published paper at a conference proceeding. The content of each chapter has been reformatted for this thesis, but is otherwise identical to the corresponding paper with the following exceptions: An errata section has been added to Chapter 5, and an error analysis to Chapter 8. Also, the camera-ready submission deadline for Chapter 9 is later than the deadline for the thesis. Minor deviations may occur.

Below is a list of the publications I have authored or co-authored during my time as a PhD student at the University of Copenhagen. The papers are written in collaboration with colleagues from the University of Copenhagen and abroad.

The two paper titles marked by $*$ are not part of this thesis.

LONG AND SHORT PAPERS

Barrett, Maria, Željko Agić, and Anders Søgaard (2015). "The Dundee Treebank." In: *The 14th International Workshop on Treebanks and Linguistic Theories (TLT)*, pp. 242–248.

Barrett, Maria, Frank Keller, and Anders Søgaard (2016). "Cross-lingual transfer of correlations between parts of speech and gaze features." In: *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pp. 1330–1339.

Barrett, Maria and Anders Søgaard (2015a). "Reading behavior predicts syntactic categories." In: *The SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pp. 345–249.

Barrett, Maria; and Anders Søgaard (2015b). "Using reading behavior to predict grammatical functions." In: *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, pp. 1–5.

Barrett, Maria, Joachim Bingel, Frank Keller, and Anders Søgaard (2016). "Weakly Supervised Part-of-speech Tagging Using Eye-tracking Data." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 579–584.

Barrett, Maria, Joachim Bingel, Nora Hollenstein, Marek Rei, and Anders Søgaard (2018a). "Sequence classification with Human At-

tention." In: *The SIGNLL Conference on Computational Natural Language Learning (CoNLL)*.

Barrett, Maria, Ana Valeria Gonzalez-Garduño, Lea Frermann, and Anders Søgaard (2018b). "Unsupervised Induction of Linguistic Categories with Records of Reading, Speaking, and Writing." In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*. Vol. 1, pp. 2028–2038.

Bingel, Joachim, Maria Barrett, and Sigrid Klerke (2018). "Predicting misreadings from gaze in children with reading difficulties." In: *Proceedings of the Thirteenth Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pp. 24–34.

Bingel, Joachim, Maria Barrett, and Anders Søgaard (2016). "Extracting token-level signals of syntactic processing from fMRI-with an application to POS induction*." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vol. 1, pp. 747–755.

Klerke, Sigrid, Sheila Castilho, Maria Barrett, and Anders Søgaard (2015). "Reading metrics for estimating task efficiency with MT output*." In: *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, pp. 6–13.

ABSTRACTS

None of the abstracts are included in this thesis.

Barrett, Maria and Anders Søgaard (2014). "Modeling eye movements when reading microblogs." In: *Association for the Advancement of Artificial Intelligence (AAAI)*. Vol. 59. 1-2, pp. 185–198.

Barrett, Maria, Joachim Bingel, Sigrid Klerke, and Laura Winther Balling (2018). "Using Naturalistic Eye-Tracking Data to Understand Children's Reading Difficulties." In: *Journal of Eye Movement Research*. Poster to be presented at *Scandinavian Workshop on Applied Eye Tracking*.

ACKNOWLEDGMENTS

I am humbly aware that one person's achievements don't grow from nothing. For me that *something* has been the open and sharing work environment of CoAStaL. A huge thank you to all present and former members of CoAStaL for contributing to a sum that is larger than all of us. Thank you for your feedback, support, and uplifting coffee break company. I am very grateful to have been part of this. A special thanks to Sigrid, Joachim, Barbara and Héctor for inspiring co-authorship and/or companionship.

The first thing I did right was to get Anders as my supervisor. I owe my mental sanity throughout my PhD to his ability to help me turn my interests into academic results. Anders is also a main driving force behind the amazing work atmosphere of CoAStaL from which I benefitted so much.

A big thank you to Frank Keller for kindly hosting me and helping me shape my project during three very productive months at the University of Edinburgh. I value your constructive feedback and the inspiring discussions.

Thank you to all my co-authors: you've all been awesome to work with. This thesis would not have been the same without each one of you.

Thanks also to the writing group and thank you Bjørn and Thomas for good company.

I'd also like to thank my friendly collegues at the Centre for Language Technologies as well as at the Department of Computer Science.

A word of gratitude to all of you who made this research possible, including my experiment participants, EyeJustRead, and the children using EyeJustRead who consented that their data could be used for research purposes.

Finally, I would like to express my sincere gratitude to Philip. I will declare my unconditional love for you elsewhere and stick to thesis matters: Thank you for your endless support, night and day shifts with our beautiful children, and homemade food. I have slept and eaten shamelessly well.

CONTENTS

I OVERVIEW AND BACKGROUND

1 OVERVIEW 3
- 1.1 Scope 4
- 1.2 Key questions 5
- 1.3 Contributions 5
- 1.4 Overview of the thesis 6
- 1.5 Errata 11

2 BACKGROUND 13
- 2.1 Where to fixate? 13
- 2.2 For how long to fixate? 15
- 2.3 General observations about eye tracking studies of syntax 16
- 2.4 Eye-tracking studies on naturalistic reading 17
- 2.5 From dependent variables in experiments to input vectors in ML models 18
- 2.6 Gaze and other human text processing metrics for NLP 20
- 2.7 Advancements in availability of eye tracking data 23
- 2.8 Theoretical definitions 24

II TWO EXPLORATORY STUDIES

3 READING BEHAVIOR PREDICTS SYNTACTIC CATEGORIES 27
- 3.1 Introduction 27
- 3.2 Experiment 28
- 3.3 Results 31
- 3.4 Related work 34
- 3.5 Conclusions 34

4 USING READING BEHAVIOR TO PREDICT GRAMMATICAL FUNCTIONS 35
- 4.1 Introduction 35
- 4.2 Eye tracking data 36
- 4.3 Learning experiments 37
- 4.4 Results 38
- 4.5 Related work 41
- 4.6 Conclusions 42

III PART-OF-SPEECH INDUCTION USING GAZE AND OTHER HUMAN TEXT PROCESSING DATA

5 THE DUNDEE TREEBANK 45
- 5.1 Introduction 45
- 5.2 The Dundee Corpus 45
- 5.3 Syntactic Annotation 46
- 5.4 Replication of Dependency Locality Theory Experiment 48

5.5 Conclusion 50
5.6 Errata 50
6 WEAKLY SUPERVISED PART-OF-SPEECH TAGGING USING EYE-TRACKING DATA 53
6.1 Introduction 53
6.2 The Dundee Treebank 56
6.3 Type-constrained second-order HMM POS tagging 57
6.4 Experiments 58
6.5 Results 60
6.6 Related Work 61
6.7 Discussion 61
6.8 Conclusions 61
7 CROSS-LINGUAL TRANSFER OF CORRELATIONS BETWEEN POS AND GAZE FEATURES 63
7.1 Introduction 63
7.2 Data preparation 64
7.3 Features 71
7.4 Experiment 72
7.5 Results 73
7.6 Error Analysis 74
7.7 Discussion 76
7.8 Conclusion 76
8 INDUCTION OF LINGUISTIC CATEGORIES WITH READING, SPEAKING, AND WRITING 77
8.1 Introduction 77
8.2 Related work 78
8.3 Modalities 80
8.4 Combining datasets 84
8.5 Experiments 85
8.6 Results 88
8.7 Analysis 89
8.8 Discussion 93
8.9 Conclusion 94

IV SEQUENCE CLASSIFICATION AND MODELLING REAL-WORLD DATA

9 SEQUENCE CLASSIFICATION WITH HUMAN ATTENTION 97
9.1 Introduction 97
9.2 Method 98
9.3 Data 101
9.4 Experiments 103
9.5 Results 105
9.6 Analysis 105
9.7 Discussion and related work 107
9.8 Conclusion 109

10 PREDICTING MISREADINGS FROM GAZE IN CHILDREN WITH READING DIFFICULTIES 111
10.1 Introduction 111
10.2 Related Work 113
10.3 Gaze Data 114
10.4 Model 120
10.5 Experiments 122
10.6 Results and Discussion 124
10.7 Conclusion 125

V CLOSING REMARKS

11 CONCLUSION AND FUTURE PERSPECTIVES 129
11.1 Limitations and challenges for future work 131

VI APPENDIX

A GAZE FEATURES 135

BIBLIOGRAPHY 137

LIST OF FIGURES

Figure 3.1 Fixation probability boxplots across five domains 28
Figure 3.2 Scatter plot of frequency and fixation probability for content words and function words. 31
Figure 3.3 Error reduction of logistic regression over a majority baseline. All domains 32
Figure 4.1 A dependency structure with average fixation duration per word. 35
Figure 4.2 Error reduction over the baseline for binary classifications of 11 most frequent dependency relations. 39
Figure 4.3 Kernel density plots across four grammatical functions of nouns. 40
Figure 5.1 An example sentence (#10) from the Dundee Corpus with UD-style syntactic dependencies and per-word fixation durations. 46
Figure 6.1 Second-order HMM. 54
Figure 6.2 Tagging accuracy on development data (token-level) as a function of number of iterations on baseline and full model. 54
Figure 7.1 Distribution of POS in the English and French training sets. 64
Figure 7.2 Two reading measures across POS class computed on the English and French training sets. 65
Figure 7.3 Accuracy on development set for all POS classes. 68
Figure 7.4 Erroneous predictions per gold POS for all combinations of training and testing language on development set. 70
Figure 7.5 Development set word type lookup in Wiktionary for English and French. 75
Figure 8.1 The percentage of overlapping word types for pairs of modalities. 80
Figure 8.2 Nearest neighbor graphs for 15 frequent nouns. 89
Figure 8.3 t-SNE plots of CCA-projected eigen dundee features for pairs of tags. 92
Figure 8.4 Learning curve assuming Wiktionary entries for k most frequent words, 93
Figure 10.1 Scanpath and fixations when reading a sentence. 112
Figure 10.2 Distributions of total number of words and misreading ratios per session after cleaning. 117
Figure 10.3 Words and misreading counts for readings of three readers in cross-user experiment 122

Figure 10.4 F_1 score distributions across test readings for each of the three readers with most sessions for three tasks. 124

LIST OF TABLES

Table 1.1 An example sentence annotated with Universal part-of-speech (UPOS) and total fixation duration in ms averaged over all readers of The Dundee Corpus. 4
Table 1.2 Overview of all human data sources used in this thesis. 8
Table 3.1 10 most used features by stability selection from logistic regression classification 30
Table 3.2 POS tagging accuracy scores on different test sets using 200 out-of-domain sentences for training. 32
Table 4.1 Most predictive features for binary classification of 11 most frequent dependency relations. 38
Table 4.2 Most predictive features for the binary classification of four most frequent dependency relations for nouns using five-fold cross validation. 41
Table 4.3 Dependency parsing results on all five test sets using 200 sentences (four domains) for training and 50 sentences (one domain) for evaluation. 41
Table 5.1 Dependency parsing results with English UD and Dundee as training sets. 47
Table 5.2 First pass durations for nouns with non-zero DLT score in the Dundee corpus. 49
Table 5.3 First pass durations for nouns with non-zero DLT score in the Dundee corpus. Corrected numbers 51
Table 6.1 Features in feature selection groups. 55
Table 6.2 Tagging accuracy on the development set (token-level) for all individual feature groups, for the best combination of groups and for the best gaze-only combination of groups. 56
Table 6.3 Tagging accuracy for the baseline, for models with no text features and for our gaze-enriched models using type and token gaze features. 57

Table 6.4 Results of an ablation study over feature groups on the test set on token-level features. 59
Table 7.1 Accuracy on development and test set for type- and token-level experiments. 67
Table 7.2 Cosine similarity between POS-averaged French and English train set gaze vectors across gaze features. Sorted by similarity. 73
Table 8.1 Results on word association norms from word-vectors.org. 81
Table 8.2 Heuristics for expanding our POS dictionary to chunks 86
Table 8.3 Chunk tagging accuracy. 87
Table 8.4 POS tagging accuracies for baselines and the model combinations that performed best on newswire development data 88
Table 8.5 Graph similarities 89
Table 8.6 Distribution of POS tags in the subset where the best model made correct predictions and the baseline made wrong predictions compared to the general development set distribution. 90
Table 9.1 Overview over the tasks and datasets used. 101
Table 9.2 Sentence classification results. 104
Table 9.3 One sentence marked as containing sexism from Waseem and Hovy (2016) development set. 106
Table 10.1 Dataset size after each cleaning step 115
Table 10.2 Overview of the feature groups used in the experiments. 119
Table 10.3 Performance across feature groups for Experiment 1. 121
Table 10.4 Statistics of (misread) words in sessions for the three readers with most readings. 123

ACRONYMS

API	application programming interface
bi-LSTM	bidirectional long short-term memory
BNC	British National Corpus
CCA	canonical correlation analysis
CPOS	coarse part-of-speech
DLT	dependency locality theory
EEG	electroencephalogram
EM	expectation maximisation
ERP	event-related potential
fMRI	functional magnetic resonance imaging
GECO	Ghent Eye-Tracking Corpus
HMM	hidden Markov model
L1	first language
L2	second language
LA	label assignment
LAS	labelled attachment score
LIX	läsbarhetsindex
MFCC	mel-frequency cepstrum coefficients
ML	machine learning
MTL	multi-task learning
MWE	multi-word expression
NLP	natural language processing
POS	part-of-speech
PTB	Penn Treebank
RNN	recurrent neural network
SHMM-ME	second-order hidden Markov model with maximum entropy emissions

SVD+IS singular value decomposition and inverted softmax feature projection

UAS unlabelled attachment score

UD Universal Dependencies

UPOS Universal part-of-speech

ZuCo Zurich Cognitive Language Processing Corpus

Part I

OVERVIEW AND BACKGROUND

1 OVERVIEW

Eye movements during reading provide one of the richest known data sources reflecting the human processing of text. It is a simple and non-invasive way of studying human cognition. During normal reading, the cognitive processing signal is implicit in the data. It has been extensively explored in psycholinguistic laboratory studies which have identified several layers of linguistic processing e.g. syntactic and morphological processing reflected in the gaze data (Frazier and Rayner, 1982; Hyönä, Bertram, and Pollatsek, 2004).

The field of natural language processing (NLP) strives to make machines capable of solving tasks that humans readily detect in the process of making sense of text. These tasks include part-of-speech (POS) tagging, parsing, anaphora resolution, sentiment classification, sarcasm detection, and semantic labeling among others. Cognitive studies have repeatedly confirmed that there is a tight relationship between word-based, eye-tracking metrics and the linguistic/lexical properties of a word. Just and Carpenter (1980) and Rayner (1977) were the first to report frequency effect and this has been confirmed many times, also when controlling for confounding factors (Rayner and Duffy, 1986). My work is motivated by the cognitive studies that report such links between cognition and gaze behavior. The key studies are presented and referenced within their individual chapters whereas Chapter 2 introduces a basic introduction to eye-tracking.

The best-performing NLP models rely on vast amounts of text that have been annotated by trained professionals[1]. These are called supervised machine learning (ML) models and learn a model based on annotated training data. Unsupervised models, on the contrary, learn patterns, for example clusters, from features in unlabeled training data. We use an unsupervised hidden Markov model (HMM) for POS induction that is type-constrained by a crowdsourced dictionary. All predictions are limited to the set of possible tags for a certain word type according to the dictionary. This makes this approach weakly supervised[2]. The model is described in detail in Chapter 6.

There are more than 7,000 languages in the world[3], but very few of them have annotated resources for training. This creates a global bias. Very few languages – and typically the major Indo-European languages – have a lot of annotated resources, but for the majority of the worlds languages, no annotated resources exist.

1 https://github.com/sebastianruder/NLP-progress
2 This distinction is used in all papers except Chapter 8.
3 http://www.ethnologue.com/world

Responsible	tourism	lies	in	absolute	respect	for	the	location	and	its	inhabitants
ADJ	NOUN	VERB	ADP	ADJ	NOUN	ADP	DET	NOUN	CONJ	PRON	NOUN
319	271	319	169	255	208	162	191	205	108	147	382

Table 1.1: An example sentence annotated with Universal part-of-speech (UPOS) and total fixation duration in ms averaged over all readers of The Dundee Corpus (Kennedy, Hill, and Pynte, 2003)

The main concern of this thesis is models not relying on annotated training data where the impact of the application seems more compelling. It is motivated by a wish to improve unsupervised NLP by using the processing signal in readers which would mainly benefit low-resource languages. The experiments have been working towards methods that do not rely on human data at test time, which is a more resource-efficient setup.

This thesis presents work that demonstrates that NLP models can benefit from including eye-tracking data from reading. We show that both supervised and weakly supervised ML models can benefit from even modest amounts of eye-tracking reading data. One study in this thesis also combines eye-tracking data with other data sources reflecting human processing, namely acoustic features and keystroke logs.

The current developments in eye-tracker hardware promise that eye-tracking data from the real world can become available on a larger scale in the very near future. This means that it will be feasible to gather larger amounts of reading data for improving NLP for low-resource languages, for example. The combination of implicitness and versatility makes eye tracking a high-gain data source for NLP.

1.1 SCOPE

My work is firmly based on the field of cognitive science, but as such, does not contribute to this particular field. It contributes to the field of NLP. Work described in this thesis was initialised by findings in the cognitive field. The studies rely on these conclusions and use machine learning (ML) and data reflecting human text processing to tackle NLP tasks.

The main data source in this thesis is word-level fixation information from eye tracking, but human text processing through keystroke logs and prosody is also explored. The gaze features represent a broad range of word-based fixation metrics. Scanpaths are not considered, apart from regressions and skips, which are considered features. The gaze feature set varies from study to study and is identified in each paper.

Language processing as reflected in direct brain measures, such as event-related potential (ERP) and measures of hemodynamic response in the brain from the field of cognitive neuroscience is outside the scope of this thesis.

POS induction is the main task of three studies of this thesis, but other studies also explore chunk induction, supervised POS tagging and supervised parsing as well as supervised detection of abusive language, sentiment classification and detection of misreadings. In contrast to the other tasks, misreading detection is not an established NLP task, but is related to complex word identification and would be useful for individual word disambiguation, as in Bingel and Søgaard (2018). In this thesis, NLP is synonymous to machine learning (ML) models, i.e. models that learn from data, rather than rules.

1.2 KEY QUESTIONS

The key question of this thesis is how data reflecting the human processing of text can benefit NLP. The answer to the following sub-questions are posed in this thesis:

- To what extent can the human processing signal for a broad range of categories be extracted from the eye movements of a reader and be used for POS tagging/parsing/POS induction?
- To what extent does the processing signal transfer from one language to another related language for POS induction?
- How will gaze data support POS induction when combined with other data sources reflecting human text processing, such as features from keystroke logs and acoustic features?
- To what extent can we use gaze features to guide the attention of a RNN for sequence classification?
- How can we model noisy real-word gaze data?

1.3 CONTRIBUTIONS

The overall contribution of this thesis is to show that human text processing metrics such as gaze, but also keystroke logs and prosody, have the potentials for improving NLP.

The studies in the thesis contribute to the field of NLP in the following way:

- Chapter 4 presents the first attempt to predict a broad set of syntactic functions from gaze and also presents the first results of improving supervised dependency parsing with gaze features.
- Chapter 3 and Chapter 6 are the first studies to improve supervised POS tagging and weakly supervised POS induction, respectively, using gaze features.

- Chapter 7 presents the first evidence that the eye-tracking signal from native reading can be used to some extent to improve POS induction on a related language.
- Although there are a couple of studies improving chunking and shallow parsing with speech and keystroke features, the study presented in Chapter 8 is the first to use keystrokes and acoustics as multi-dimensional, continuous features for POS and syntactic chunk induction. It is also the first to combine them with other modalities (gaze and pre-trained word embeddings) for POS induction and chunk induction.
- Chapter 9 introduces a novel way of regularizing the attention of a recurrent neural network (RNN) with human attention derived from continuous gaze features. We show that we can use the inductive bias from human attention to improve performance on several NLP tasks, such as detection of abusive language, grammatical error detection, and sentiment classification.
- Chapter 10 presents the first study to model real-world, eye-tracking reading data from children's reading sessions. In this study, we predict misreadings using gaze features from reading.

1.4 OVERVIEW OF THE THESIS

This section will explain how the articles of this thesis are connected. References are kept at a minimum for readability in this section since each chapter contains its own references.

1.4.1 *Two pilot studies*

The studies in Part ii represent first evidence that the signal in gaze features is usable for cross-domain POS tagging and dependency parsing, respectively. Both studies use eye-tacking data that is collected for this purpose.

Two hundred and fifty English sentences from established corpora, randomly sampled from five different domains were read by 10 native speakers. Chapter 3 contains the methodological description of the data collection.

The notion of *domain* is frequently used in NLP, although it is not clearly defined. It usually refers to a collection of texts of the same genre or source from the assumption that this common denominator will have some systematic impact on the vocabulary or writing style. But there are many underlying factors that are ignored in this process. See Plank (2016b) for a discussion.

In Chapter 3, we explore the well-established fact from cognitive studies, that readers are less likely to fixate their gaze on closed-class

syntactic categories, such as prepositions and pronouns and more likely to focus on words belonging to open-class categories, such as verbs and nouns. But we go a step further and try to investigate to what extent the coarse-grained syntactic category of a word in context can be predicted from gaze features from reading. Our results show that gaze features do discriminate between most pairs of syntactic categories, and we show how we can use this information to tag words with POS across domains, when tag dictionaries enable us to narrow down the set of potential categories. We also show that we can improve supervised POS tagging on all domains. The improvement is even larger when we also add type-level gaze features from the Dundee Corpus and word length and word frequency features.

Since we can predict word class, we take the investigation a step further in Chapter 4. Here we investigate the more fine-grained distinctions between grammatical function e.g. the subject or object function for nouns. In addition, we show that gaze features can be used to improve a discriminative, transition-based dependency parser. We can improve dependency parsing over several baselines across domains, e.g. pre-trained word embeddings.

Both exploratory supervised experiments show, that even though there is some variation across the five domains, we can successfully predict grammatical functions and grammatical categories across domains. Both papers explored predictive eye-tracking features and they found overlapping features for POS and dependency relations. To continue this line of work, it was crucial to obtain more syntactically annotated gaze data.

1.4.2 *More gaze data - and other human resources used in the remaining articles*

Table 1.2 present an overview of all human sources used in this thesis. Note that there is also a task-solving part of the ZuCo corpus, a Dutch and L2 English part of the GECO corpus, and a transcribing part of the keystroke dataset. They are not included in the table, since they are not included in this thesis.

Only a couple of years ago, the Dundee Corpus (Kennedy, Hill, and Pynte, 2003) was the largest eye-tracking corpus by token count. But it was not syntactically annotated. Because both pilot studies showed, that we could predict syntactic categories and functions cross-domain, the easiest way to get more data was to obtain syntactic annotation the Dundee Corpus. It contains around 50,000 tokens for English and French. It is read by ten subjects in each language and is extensively used in larger-scale cognitive science studies. It contains naturalistic reading of contextualized running text. The subjects read articles from either *Le Monde* or *The Independent*. We made a Universal Dependencies (UD)-style annotation (Agić et al., 2015) for the English

Name	Ref	Lang	Modality	Tokens	Types	*n* subj	Available annotations
250 cross-domain	Barrett and Søgaard (2015a)	en	gaze	3,241	1,372	10	According to sources
Dundee	Kennedy, Hill, and Pynte (2003)	en	gaze	51,502	9,776	10	UD
Dundee		fr	gaze	47,445	12,464	10	French Treebank 1.4
EyeJustRead	not released	da	gaze	8,681	1,539	44	Misreadings
Free composition	Killourhy and Maxion (2012)	en	keystroke	14,890	2,198	20	
GECO	Cop et al. (2017)	en	gaze	54,364	5,817	14	MWE
Prosody	Frermann and Frank (2017)	en	prosody	763,854	598	22	
ZuCo	Hollenstein et al. (2018)	en	gaze	20,765	4,651	12	Relations and sentiment

Table 1.2: Overview of all human data sources used in this thesis. For EyeJustRead, counts are on the cleaned dataset. Note that for EyeJustRead data, un-fixated words are not in the corpus opposed to the other eye-tracking datasets. Numbers are therefore not completely comparable to this dataset. In all corpora containing punctuation, the tokenisation follows visual units. Token count is counted on this tokenisation, whereas the type count is counted on words stripped from punctuation at the end or beginning of the token (but for English and French, contractions are preserved). For the French Dundee, numbers are on the entire corpus, though 5% of the words are not found in the French Treebank and therefore not included in Chapter 7. The French Dundee corpus contain 52,173 tokens per participant and 11,321 types according to Pynte and Kennedy (2006). The difference could be due to another tokenisation than visual units before counting.

part. This annotation work is described in Chapter 5 and is freely available[4]. The Dundee Corpus is used in all of the following articles except Chapter 10.

To obtain syntax annotations for the French part of the Dundee Corpus for Chapter 7, the French part of the Dundee Corpus was manually merged with the syntactic annotation of the French Treebank. The read texts originated from this source. This merge is not publicly available due to licensing restrictions. 2,518 tokens could not be merged, because they could not be located in the French Treebank.

In Chapter 8, we also include the newly released Ghent Eye-Tracking Corpus (GECO). It is now considered the largest, publicly available eye tracking corpus. The monolingual part is slightly bigger than the English Dundee Corpus both with respect to number of participants and token count (14 subjects read over 54,000 tokens). It contains first language (L1) (English and Dutch) and second language (L2) (English) reading of an entire novel. We only used the English L1 part. It is not syntactically annotated, which was not needed for Chapter 8.

In Chapter 8, we also used keystroke logs from a typing experiment and prosody features from a large corpus of child-directed speech. Both data sources are only used in this study and described more closely in the respective chapter.

In Chapter 9, we include a newly released English resource: the ZuCo Corpus. The normal reading part of the ZuCo Corpus contains

4 https://bitbucket.org/lowlands/release/src

eye-tracking data (as well as EEG that is not used in this thesis) for 300 relation-annotated sentences and 400 sentiment-annotated sentences.

Chapter 10 is the only study modelling real-world reading data recorded with a consumer-grade eye tracker. The data comes from real reading sessions by Danish children with reading difficulties. It is annotated for misreadings by the teacher. The data poses other challenges with respect to pre-processing as described in detail in the chapter.

For all data sources reflecting human text processing only naturalistic text processing is included. Experimental factors or tasks would affect the human text processing. Only keystroke logs from the free composition condition and ZuCo gaze data from the normal reading conditions are used. Unbiased human data does not exist, and even naturalistic data is biased. Therefore, the Dundee and Ghent Eye-Tracking Corpus (GECO) corpus are treated as two separate modalities in Chapter 8, and the ZuCo Corpus and the Dundee Corpus are normalized separately in Chapter 9.

1.4.3 *Three studies on weakly supervised POS induction*

Since supervised POS tagging is practically a solved task (see Manning (2011) for a discussion) the following work moved to weakly POS induction using gaze and other human text processing data. For many of the world's languages, there are no or very few linguistically annotated resources. But raw text and sometimes also dictionaries, can be harvested from the web. These resources can be used to train weakly supervised POS taggers.

Since my previous work showed that gaze can be used to discriminate POS signal, we perform POS induction experiments using this and other human data. The following experiments are on either English or English / French, due to the availability of eye-tracking data in these particular languages. For the following three experiments, we train a weakly supervised POS tagger using a second-order hidden Markov model with maximum entropy emissions (SHMM-ME). Its predictions are constrained by a crowd-sourced dictionary.

The best model in Chapter 6 uses type-level aggregates of English eye-tracking data and significantly outperforms a baseline that does not have access to eye-tracking data. This model also outperforms models using token-level features. The best model includes all gaze features. Type-level gaze features further have the added positive benefit of not requiring gaze at test time.

Both pilot studies in Part ii explore a broad set of word-level gaze features. Results indicate that the signal is distributed over several features. Chapter 6 explore these features in groups of 3-6 and find that the best model indeed includes all features. Even the best indi-

vidual gaze feature group is beaten by frequency and word length features.

In Chapter 7, we move further and show that gaze and POS correlations largely transfer across the related languages English and French. This means that we can replicate the previous English study on gaze-based POS induction, but now for French. We find that we can use English gaze data to assist the induction of French POS. Type-level features are also here consistently better than token-level. The following experiment therefore only use type-level-aggregated features of human text processing.

In Chapter 8, we explore the performance of the model in a different direction: by including several other data sources, presumably containing human syntactic processing signal. We find that performance can be improved by combinations of only partially overlapping resources from eye-tracking, prosody, keystroke, and pre-trained word embeddings. The best models significantly outperform a baseline of pre-trained word embeddings. Our analysis shows that improvements are even bigger when the available tag dictionaries are smaller. Based on the findings in Chapter 7 and Chapter 6, we only use type-level averaged features in this study, and we are therefore able to evaluate our models on established NLP corpora.

A fourth POS induction study I co-authored is not included in my thesis but has several methodological similarities to the three studies above. In Bingel, Barrett, and Søgaard (2016), we used the fMRI signal to improve POS induction. The fMRI-signal posed other preprocessing challenges as well as theoretical challenges and the signal was less reliable than gaze.

1.4.4 *Sequence classification and modelling real-world data*

Chapter 9 is exploratory with respect to how the gaze is included in the model and it is the only study in this thesis concerned with sequence classification. Opposed to the previous studies, the eye-tracking feature is not used to directly reflecting the human processing of the NLP task. Rather, the gaze is used to regularise the attention of the supervised model. Learning attention functions for recurrent neural networks requires large volumes of data, but many NLP tasks simulate human behavior. In Chapter 9, we present a RNN architecture that jointly learns the recurrent parameters and the attention function, but is able to alternate between supervision signals from labeled sequences and from attention trajectories in eye-tracking data. We show substantial improvements across a range of tasks, including sentiment analysis, grammatical error detection, and detection of abusive language.

Chapter 10 is facilitated by the cheaper and available eye trackers. Data comes from real reading class sessions and is captured by a

consumer-grade eye tracker. We try to predict where the child misreads a word, which would be useful for, for example, personalized reading assistance and an important tool for the reading teacher. We use an ensemble of neural models and decision trees for experiments across the entire dataset and a multi-task learning setup to explore whether predictions generalize across readers. Our experiments show that despite the noise and small number of misreadings, gaze data improves the performance more than any other feature group and achieves good performance. We further show that gaze patterns for misread words do not fully generalize across readers, but in some cases, we can transfer knowledge between readers using multi-task learning.

1.5 ERRATA

After publication, I was made aware that my calculation of the integration cost for the reproduction of the (dependency locality theory (DLT)) experiment for Chapter 5 contained a systematic error. The DLT score is only used in this particular study, to show one application of the new syntactic annotation of the Dundee Corpus. I am grateful to Scarlett Hao for pointing this out. I erroneously assigned the integration cost for the number of intervening new discourse referents to the head instead of to the dependent.

When re-running the DLT experiment, I became aware that the calculation of the gaze feature, the First Pass Duration feature on the Dundee Corpus was erroneously extracted by my script for around 5% of the words, namely the words that were re-fixated. The Dundee Corpus does not come with ready-to-use gaze features (like for instance the GECO corpus). Instead, the Dundee Corpus provides raw gaze data, which each researcher can use to extract the desired features. The First Pass Duration values were higher than they should have been due to the re-fixation(s) being added twice to the First Pass Duration, instead of only once. It was systematic for all words being re-fixated in the first pass. The bug only affects one gaze feature and only 5% of the words. I fixed the bug and re-extracted the feature from the raw Dundee Corpus.

The study, most affected by the bug in the extraction of the First Pass Duration is also the only study affected by the bug in the calculation of the DLT integration score, presented in Chapter 5. Here I perform a statistic test only on the First Pass Duration using DLT as a fixed effect. I re-ran this test using the corrected scores for DLT and First Pass Duration and present the new results in the last section of the relevant chapter.

The First Pass Duration is used in several studies along with around 30 other gaze features from the Dundee Corpus. I did not rerun these

experiments using the corrected First Pass Duration score. I do not expect this to have any significant impact on these studies.

2 BACKGROUND

Reading is a complex interplay of several processes not yet fully understood. Some processes are low-level oculomotor factors (typically associated with where to look) and others are higher-level lexical properties (typically associated with the duration of looking). Context factors (such as predictability from preceding words) also play a role. This section presents a background overview from psycholinguistics about well-established processes that guide *when* and *where* the eyes move during adult skilled reading. It will introduce the basics of some of the different levels of text processing to give the reader an understanding of what can be expected with respect to the high-level processing signal that the studies include. Uncited numbers and facts in this section come from Juhasz and Pollatsek (2011), Rayner (1998), and Staub and Rayner (2007) where further references can be found. Specific, high-level effects relevant to each study, can be found in the respective chapters.

Unlike ERP (such as EEG) and measures of hemodynamic brain response (such as fMRI), eye tracking measures are proxies for the cognitive processes, whereas brain-metrics may be considered direct measures of cognitive processes. Eye-tracking data contains processing information with high temporal resolution, but does not provide information about *which* process occurs. Since many layers of human text processing occur and co-occur, eye-tracking during reading is usually studied in controlled experiments where only the parameter of interest varies and any variation on gaze behaviour can be attributed to this change. But for NLP we need to generalize to unseen text and are thus interested in naturalistic reading of representative text.

2.1 WHERE TO FIXATE?

Contrary to the internally perceived experience of reading, the eyes do not glide smoothly over the text, but instead perform a series of rapid, ballistic movements called saccades. Only a small part of the visual field can be seen sharply by the eye at a time, and the main function of the saccades is to bring another part of the visual field in focus. Saccades last for about 20-40 ms on average and most often move the eyes 7-9 letter spaces forward. But around 10% of the saccades – often unconscious to the reader – go backwards. Backward saccades are referred to as regressions. Between saccades, the position of the eye is relatively stable for 50-500 ms, but typically 2-300

ms. These are called fixations and it is only during fixations that the eye is still enough for the brain to be able to perceive any information.

Due to the anatomy of the eye, only the foveal area, which is 1°of the visual field is seen very sharply during a fixation. The parafoveal area covers around 5°of the visual field around a fixation. The acuity drops gradually and quickly. Beyond that, in the peripheral vision, the reader is aware of the general shape of the text e.g. line breaks.

In combination with the anatomic foveal limitations, there seems to be an unconsciously learned aspect: For a left-to-right reader, the viewed text part during reading is not symmetrical. It extends 3-4 characters to the left and 14-15 characters to the right of a fixation (McConkie and Rayner, 1976; Rayner, Well, and Pollatsek, 1980). The area where words can be identified, *the perceptual span*, is even smaller and only extends 7-8 characters to the right and varies as a function of text difficulty. The asymmetry of the perceptual span for right-to-left readers is mirrored (i.e. shifted to the left) compared to left-to-right readers. It is even found that bi-lingual readers of Hebrew and English unconsciously mirror their perceptual span to suit the reading direction (Pollatsek et al., 1981). The perceptual span is important, especially when examining *where* the eyes move and trying to understand the nature of preview effects. The differences across writing systems are important to keep in mind when discussing the multilingual perspectives in the conclusions of this thesis.

From the length of the typical saccadic span (7-9 letter spaces) and the size of the perceptual span (around 12 characters), it is obvious that some words are not fixated at all, but are visible to the reader during a fixation on the previous word. But preview effects occur in a complex interplay with predictability and frequency effects. The perceptual span is part of the explanation why not all words are fixated and an even bigger part of the explanation for short words. Readers generally benefit from previewing, and several studies have explored the characteristics of the information the reader obtains from the parafoveal view. Readers seem to have access to the letters and the sound codes of the word but not the the meaning of the word nor the morphological composition before the word is fixated. Words are also more often skipped when they are predictable from context or very frequent. Skipping words can be understood as the consequence of a successful preview processing (Pynte and Kennedy, 2007). The perceptual span can, to some extent, explain why Carpenter and Just (1983) found that 38% of function words are fixated and 83% of content words are fixated, since function words are generally shorter than content words and can thus more often be previewed. But when controlling for word length, function words were still skipped more often. Word length is also robustly positively correlated with fixation probability, number of fixations as well as fixation duration. Staub and Rayner (2007) conclude that low-level visual information is most

important with respect to *where* to move the eyes, whereas linguistic factors seem more important with respect to *when* to move the eyes. This thesis therefore focuses mainly on fixation durations.

An interesting feature of eye-tracking data is that it allows us to study regressions. In normal skilled reading around 10–15% of all saccades are regressions, most of them short regressions. Regressions longer than 13 letter spaces are rare. In the review of Rayner (1998) it is concluded that regressions are not fully understood, but they are positively correlated with text difficulty; they may come from comprehension difficulties, but also occulumotor errors; and they are also linked to recovering from syntactic parsing errors (Frazier and Rayner, 1982).

2.2 FOR HOW LONG TO FIXATE?

The two most well-researched properties that guide how long a word is fixated, are word length and word frequency. People look longer at long and infrequent words. But there are many other properties proven to affect fixation duration: Morphology, orthography, familiarity, number of meanings, and estimated age of acquisition are shown to have frequency-independent effects. Relational effects, such as predictability from context, are also well studied. Word length and word frequency account for more of the variance in mean fixation duration than nine other known factors (high- and low level) combined. In the case of this thesis, where we do not try to understand human cognition but rather try to extract the syntactic processing signal, it is important to keep in mind that the signals of interest will be mixed with other processing signals. The intrinsic properties of a word are constant for an unambiguous word, but may interact with contextual features.

One of the most intriguing features of eye tracking is that it allows us to study early and later text processing separately. The early measures (e.g., first fixation duration and first pass duration) have been shown to be sensitive to early text comprehension processes, such as lexical access and early syntactic and semantic integration. In contrast, late measures (e.g., total fixation duration and information about later passes over the text) are sensitive to later cognitive processes, such as information reanalysis, discourse integration, and recovery from processing difficulties.

Because numerous studies (Juhasz and Rayner, 2003; Rayner, 1977; Rayner and Frazier, 1989) have found that fixation duration is sensitive to various lexical properties of the fixated word, we can assume that there is a rather tight relationship between what is fixated and the cognitive processes on the word level, what Just and Carpenter (1980) referred to as the eye-mind span. Preview effects and spillover effects (the processing of a word results in extended fixation duration

on the next word) are examples of phenomena where this relation is not air-tight. Word frequency affects first fixation duration on the fixated word, but is also an example of a phenomenon, that is known to cause spillover effects to the following word (Rayner and Duffy, 1986). It is therefore sensible to provide information about fixations and some properties on the surrounding words to ML models.

Processing effects found on super-word level (e.g. multi-word expression (MWE)) and sub-word level (e.g. morphology) are other exceptions. In behavioral experiments, including eye tracking experiments, the entire MWE is found to have a processing advantage over novel strings of language. Siyanova-Chanturia (2013) provide a review of eye-tracking evidence. For sub-word level processing, e.g. Hyönä, Bertram, and Pollatsek (2004) find that the frequency of the first morpheme, and to some extent of the second, influences fixation duration, suggesting that words are decomposed into morphemes as they are analyzed. But the vast majority of studies show word-level effects.

2.3 GENERAL OBSERVATIONS ABOUT EYE TRACKING STUDIES OF SYNTAX

There are a lot of psycholinguistic studies which focus on the processing of syntax, but most studies are interested in parsing of temporarily ambiguous syntax. Watching humans resolve temporarily ambiguous syntax is informative about human cognition, but the phenomenon only applies to a small subset of naturally occurring sentences. In an extensive and exhaustive survey of sentence comprehension factors, approximately 70 studies out of 100, all studying higher-level effects, deal with temporarily ambiguous syntax/garden path sentences. Other high-level effects within the scope of the survey are semantic, pragmatic and world-knowledge factors (Clifton, Staub, and Rayner, 2007). One take-home message is that human syntactic parsing is sequential and incremental and probably not sensitive to the hierarchical structure (Frank and Bod, 2011; Frazier and Rayner, 1982) (though see Fossum and Levy (2012) for a discussion).

There are examples of studies aiming for a broad-coverage model of human syntax processing, e.g. Demberg and Keller (2008) where the effects of delexicalized surprisal (by POS) and dependency locality theory (DLT) on fixation durations are explored. Somewhat related to Demberg and Keller (2008), Pynte, New, and Kennedy (2009) explore semantic processing as an effect of priming on the French Dundee Corpus for nouns, verbs and adjectives. In this study, the primes are the previous content words and the study differentiates between remote and local priming. Their results suggest that verbs elicit remote effects and nouns and adjectives elicit local priming effects. Schmauder, Morris, and Poynor (2000) find frequency effects

and word length effects in the processing of both content and function words as well as increased processing time in the phrase immediately following a low-frequency function word.

Very little research is dedicated to broad-coverage studies of word classes, besides the coarse-grained distinguishing between content and function words. Pynte and Kennedy (2007) study the effect of punctuation on the English Dundee Corpus and explore word class as a confounding factor. They find that the more probable the word class of a word, the higher the skipping probability of this word. They also find that punctuation, which is a strong indicator of the word class of surrounding words (e.g. the punctuated word is most often a noun and the next word is most often a closed class word), does not in itself triggers word skipping, but the punctuation is a further cue to the word class prediction and thus triggers skipping. For example, a punctuation changes the probability that the previewed word will be a function word. Supporting this finding, both Bauman (2013) and Demberg and Keller (2008) find that the probability of a word class is negatively correlated with of fixation duration in naturally occurring text. Distinguishing or characterizing word classes by gaze behaviour is outside the scope of all the above studies, but all studies reveal that word class processing is also dependent on the context. Only Furtner, Rauthmann, and Sachse (2009) find that nouns in text with jumbled letters are regressed to more often than any other word class. This also seems to be the case for L2 reading (Furtner, Rauthmann, and Sachse, 2011). The only study is Zelenina (2014) where the output of a POS tagger was re-ranked based on eye-tracking features, but this work is, however, not published beyond the master thesis.

No cognitive studies explore a full set of word class differences. It may be that the difference in processing word classes is not a well-motivated cognitive task, whereas the grammar-lexicon distinction (roughly content words vs. function words) is fuelled e.g. by findings of agrammatism in some aphasia patients[1] and battles between linguistic theories. Another explanation is that the differences are most likely reflected in a complex interplay between many gaze features. Cognitive studies usually analyse one gaze variable at a time, e.g. in a factorized experimental design using statistical tests. It is quite unlikely that one gaze variable is able to distinguish a broad set of word classes.

2.4 EYE-TRACKING STUDIES ON NATURALISTIC READING

In order to include gaze in NLP, the processing signal found in controlled experiments needs to transfer to naturalistic reading of larger

1 though there is also evidence that e.g. nouns and verbs are processed differently in the brain, e.g. (Luzzatti et al., 2002)

quantities of naturally occurring text. This section will present evidence that this is the case.

Reader models, such as the E-Z Reader model (Reichle et al., 1998), provides a model for when and where the eyes move during reading based on word identification, visual processing, attention, and oculomotor control. Some studies have tried to model aspects of global reading strategies, e.g. Hara, Kano, and Aizawa (2012) and Nilsson and Nivre (2009). Matthies and Søgaard (2013) even show that word-skipping behaviour generalizes across readers.

A related line of research shows how conclusions from psycholinguistic experiments generalize to larger sets of naturalistic reading but without a specific NLP focus. Many use the Dundee Corpus as a data source. This research covers uncertainty of word identity (Frank, 2010; Hahn and Keller, 2016; Smith and Levy, 2010), surprisal and other expectation-based resource allocation (Frank et al., 2013b; Frank and Thompson, 2012; Hale, 2001; Levy, 2008; Monsalve, Frank, and Vigliocco, 2012; Rauzy and Blache, 2012; Shain et al., 2016b), parafoveal effects on foveal reading time (Kennedy and Pynte, 2005), frequency and predictability effects (Kennedy et al., 2013), comparison of global reading patterns for English and French (Pynte and Kennedy, 2006), and broad-coverage theories of syntactic processing, such as DLT and delexicalized surprisal (Demberg and Keller, 2008).

Furthermore the larger amount of data allows such studies to model continuous functions of, for example, surprisal and frequency as opposed to the discretised categories of psycholinguistic experiments. My work is motivated by the fact that these studies find evidence of global psycholinguistic conclusions on corpus data. The processing of syntax and word classes on a broad set of categories has not been sufficiently covered by the psycholinguistic field, however, so my work should be considered exploratory. Moreover I try to apply human text processing data to solve existing NLP tasks.

2.5 FROM DEPENDENT VARIABLES IN EXPERIMENTS TO INPUT VECTORS IN ML MODELS

NLP as engineering is not theoretically accountable to the same extent as cognitive science studies when it does not claim to explain why models work. Rather NLP sets up external modes of evaluating model robustness. This thesis belongs to the field of NLP. This section will try to explain in words the journey of the features from dependent variables in cognitive studies to parts of input vectors in ML models.

In controlled cognitive experiments with limited textual data, the tested parameter is often discretised (e.g. high vs. low frequent words). With more data, it is possible to capture an effect over a continuous spectrum. All studies in this thesis use continuous values. This is not

dissimilar to what cognitive studies would do, if more textual data was available per experiment.

Although psycholinguistic eye-tracking studies use a large number of eye-tracking metrics, each study usually only analyses a couple of pre-defined and theoretically motivated metrics. In his extensive review, Rayner concludes that "It thus appears that any single measure of processing time per word is a pale reflection of the reality of cognitive processing."(Rayner, 1998, p.377). When we are not interested in isolating one effect and making conclusions about human cognition, but rather improve NLP models (the engineering approach), it seems rational to include a richer set of eye-tracking features, i.e. multidimensional gaze representation. All studies in this thesis except Chapter 9, use multidimensional, continuous features. In Chapter 9, we use one continuous feature.

If not controlled for, some extraneous variables will distort the results of cognitive studies or in the worst case scenario: be confounding. In controlled studies, some variables are controlled for by sampling or by being kept constant. These include word length, word frequency, predictability, age of acquisition, sentence length, position on screen, position in sentence, frequency of next word, and/or frequency of the previous word. In NLP, some factors are easy to control for by just adding them as features. The ML algorithm will pick up the effects and interaction effects. Carpenter and Just (1983) found in a linear regression model that word length and word frequency together account for 69% of the variance in mean fixation duration when reading scientific texts, whereas nine other factors together only accounted for 37% of the variance. The factors are intercorrelated meaning that the numbers are not additive. The nine other factors coded e.g. whether the word was at the beginning of a line, at end of sentence, was the first content word in passage. All 11 variables together accounted for 79% of the variance. In controlled studies, such factors are controlled for by only studying the relevant effect on target words of the same length and from the same frequency group. It is worth taking the most robust features into account to ensure that our model is not swallowed by noise, but also balance the gain against the feature engineering effort. Furthermore, basic features can be used as baselines when evaluating the models.

On a related note, it is also possible to control for e.g. spillover effects by providing the frequency of the previous word and the fixation duration of the previous word as features. Likewise we can control for preview effects by providing some features of the upcoming word. The success of all control approaches is dependent on the selection of the right features.

Overfitting is a risk in all NLP studies. When using relatively little human data from a small number of subjects, it represents real threat. The token-level features should not be overly exposed to the risk of

overfitting, but word-type-averaged features are. The equivalent for cognitive science studies would be experiment bias. It is a living condition that can be minimized but must otherwise be documented and accepted. An engineering approach would be to obtain more data for future experiments. This is also discussed in Chapter 11.

2.6 GAZE AND OTHER HUMAN TEXT PROCESSING METRICS FOR NLP

This section aims to provide a unified review of NLP tasks[2] for written text processing informed by human text processing measures from reading. The source is primarily eye-tracking data, but also keystroke, acoustic cues and brain imaging are included when available. Studies modelling human data without an application to NLP is not included.

Human processing data is a fairly new addition to the field of NLP, and this section mainly aims to show the width of the subfield. Although the list of studies is not exhaustive, the tasks and applications are intended to be. I also include a few relevant studies, where gaze is not directly used to solve NLP tasks for text, but where the methods used or the conclusions could be adapted to NLP with few or no modifications. That is, all studies below are included because they could be used to generalize to unseen text and evaluated for some NLP task. The studies are limited to models that apply to all words or at least a broad range of words. This section is structured thematically in an attempt to provide an overview of an emerging field. Studies into the cognitive processes of annotators are not included since this is a different cognitive task from pure reading. But I include studies that evaluate the output of NLP models using implicit human text processing.

A well-researched line of work shows, that eye-tracking data from reading reflects text comprehension, e.g. Green (2014) and Singh et al. (2016). This is an obvious direction to pursue since processing difficulties are expected to show larger effects on fixation durations as well as regressions. Text comprehension of individual readers is considered peripheral to the NLP field. However, I only include studies that deal with predictions/scoring on natural reading. Thus, all methods/scores could be used by ML models and evaluated. In von der Malsburg and Vasishth (2011) and von der Malsburg, Kliegl, and Vasishth (2015), a sentence-level scanpath score that can be used to measure irregularities in reading e.g. by detecting syntactically complex sentences is presented. It is evaluated on laboratory reading data but it is in theory applicable to real-world data. Similarly, Wallot et al. (2012, 2014, 2015) score the reading fluency of text passages based on different metrics of scanpath regularity and use these scores to predict reading comprehension.

2 This is not a completely closed set of tasks

The following studies are examples of using the comprehension signal in eye movements with a tighter NLP relation. Klerke, Goldberg, and Søgaard (2016) predict discretised gaze features as an auxiliary task in a bi-LSTM to improve sentence compression. They employ an early gaze measure (First pass duration) and a late measure (regressions), individually.

Eye movements have been used to evaluate the quality of NLP model output. If eye trackers are being built into mass market products, such as smartphones, tablets and computers it is not difficult to imagine the benefit of having ML output evaluated during use without the need for any explicit feedback. Gaze has been used to directly evaluate the quality of machine translation output. In Klerke et al. (2015), different versions of logic text puzzles were solved by native-speaking human participants. Both correctness of the solutions and eye-tracking metrics were used to compare automatically translated versions to expert translated versions. Reading metrics were better proxies for the usability of the text than the standard, automatic metric: bilingual evaluation understudy score (BLEU) (Papineni et al., 2002). Stymne et al. (2012) also evaluate machine translation output using eye tracking and comprehension question, but did not include a performance metric such as puzzle solving. They also found longer gaze time on machine translation error. Eye tracking has also been used to evaluate automatically compressed sentences (Klerke, Alonso, and Søgaard, 2015), translation difficulty (Mishra, Bhattacharyya, and Carl, 2013), and – together with fMRI – word embedding quality (Søgaard, 2016).

Another line of research extracts the semantic processing from data reflecting human text processing. Understanding sarcasm during reading requires extra processing time, which is also detectable from the eye movements. *The Sarcasm Processing Time* depends on the degree of context incongruity between the statement and the context (Ivanko and Pexman, 2003). Mishra, Kanojia, and Bhattacharyya (2016) and Mishra et al. (2016a) use this phenomenon for supervised sarcasm detection. Eye-tracking data has also been successfully used for sentiment classification (Mishra et al., 2016b) and to evaluate the relevance for information retrieval purposes, e.g. Hardoon et al. (2007) and Salojärvi et al. (2003). Loboda, Brusilovsky, and Brunstein (2011) also show that relevance can be inferred from gaze at word level. Especially word-level relevance plays a role for guiding attention functions of RNNs. There are also attempts with named-entity recognition with corpus eye-tracking data (Rotsztejn, 2018). This study show that there is a significant, negative correlation between the number of times an entity is mentioned and reading time. This is also validated on a held-out test set.

MWEs cover a heterogeneous group of expressions. They vary to a in their linguistic properties but they are perceived as highly con-

ventional by L1 speakers (Siyanova-Chanturia, 2013). There are some attempts to predict MWE using eye-tracking data (Rohanian et al., 2017). Yaneva, Taslimipoor, Rohanian, et al. (2017) compare gaze patterns from the large GECO corpus of naturalistic reading annotated for MWEs. They compare verb-particle and verb-noun MWEs to control phrases of the same POS pattern. They find that MWEs are generally processed with lower numbers of fixations and lower reading times than control sentences, but that early gaze measures do not discriminate MWEs from novel, matched strings. They also find that there is a processing advantage for the last word of the MWE, if it is a noun. This applies only to L1 speakers, not L2. Particles occurring at the end of a sentence were processed equally efficiently for MWE and control phrases. Although the difference is subtle, this study is interesting and important because it confirms that in naturalistic reading contexts, word processing also occurs at super-word level, and that gaze can be used to discriminate formulaic language from novel phrases.

There are a couple of supervised parsing experiments using human data. Plank (2016a) used a single, discretised keystroke feature, pre-word pause from keystroke logs, to help supervised, shallow parsing. It builds on the intuition that pre-word pauses are longer before syntactic chunks, so words belonging to the same chunk are typed in bursts. The keystroke feature was an auxiliary task in a multi-task bi-LSTM model. Using keystroke improved two shallow parsing tasks: CCG supertagging and chunking over text annotations alone. The following two studies also used a single, discretised feature: Pate and Goldwater (2011) used a single acoustic feature to aid unsupervised chunking and later also parsing (Pate and Goldwater, 2013). Like Gonzalez-Garduno and Søgaard (2018), we predict one continuous feature. Other relations between words can also be established. Jaffe, Shain, and Schuler (2018) use gaze to improve co-reference resolution.

Some models include what could be referred to as second order human processing i.e. include metrics motivated by the fact that they are known to correlate with human text processing or mimic human behavior. The following are examples of this: Shain et al. (2016c) include human language acquisition features in an unsupervised parser: limited working memory capacity as well as a human left-corner parsing strategy. Wang, Zhang, and Zong (2017) include word-level features, that are known to correlate with human attention (e.g. surprisal) to weigh the attention for sentence representation.

Other studies, in the periphery of the scope of this thesis, are concerned with modelling characteristics of the reader. Berzak et al. (2017) predict the native language of the reader based on eye-tracking data of L2 English reading. Augereau et al. (2016) and Berzak, Katz, and Levy (2018) use gaze data to predict English proficiency from gaze data.

2.7 ADVANCEMENTS IN AVAILABILITY OF EYE TRACKING DATA

Eye-tracking research has a more than 100 year-old history, but for many years, eye trackers were unica machines or very expensive and only found in big research labs. In the past ten years, there have been several attempts to make eye-tracking available through a regular webcam (e.g. San Agustin et al. (2009)) and smartphone (Krafka et al., 2016), just to mention a few. In 2013, Samsung Galaxy S4 came with a built-in eye tracker and the following year, the first $100 eye tracker was available. Since then, more hardware options have emerged, also from established eye-tracker companies. A more recent advancement is iPhone X which has API access to all necessary components for external developers to implement eye tracking.

Low-cost gaze technologies have taken big leaps during the course of my PhD, meaning that towards the end of my PhD, we have been able to model real-world gaze data from children's reading sessions captured by a Tobii eye tracker targeted at regular consumers. Such data was not available four years ago. The data was kindly provided – with consent from all participants and parents – by a Danish startup company, which is one of the first to bring real-world, eye-tracking data from reading into practical use. Another example on an application benefiting from the increased availability of eye trackers include iAppraise (Abdelali, Durrani, and Guzmán, 2016), which is a machine translation evaluation environment supporting eye tracking using a low-cost eye tracker.

Recently there has been an increased interest in naturalistic eye-tracking data. When I started my PhD, the only large corpus of naturalistic reading was the Dundee Corpus (Kennedy, Hill, and Pynte, 2003). Both parts contain contextualized reading of naturally occurring text, as opposed to the Potsdam Sentence Corpus (Kliegl et al., 2004), which comprises 144 constructed, German sentences read by 222 participants. Other databases include Frank et al. (2013a) which consists of 205 naturally occurring English sentences read by 43 native speakers, and the DEMONIC database Kuperman et al. (2010) which consists of 224 constructed Dutch sentences read by 55 participants.

However, more resources that may accommodate NLP have recently emerged. The GECO Corpus (Cop et al., 2017), described in Section 1.4.2 is now the largest eye tracking corpus by token count. Like the Dundee Corpus, it contains naturally-occurring contextualized text. The Provo Corpus (Luke and Christianson, 2018) contains around 2,700 English words read in 55 short paragraphs by 89 native participants. The Provo corpus comes with predictability norms. The ZuCo Corpus (Hollenstein et al., 2018), also described in Section 1.4.2, contains gaze and EEG data from reading isolated sentences. The naturalistic reading parts contain 700 annotated sentences (sentiment or relations).

There is reason to believe that eye tracking technology will soon be available on a larger scale capable of providing data that can be used for NLP. This expectation calls for ways to benefit from this rich data source. The following studies present first evidence that NLP should consider including eye-tracking data and other rich data sources reflecting human processing of text.

2.8 THEORETICAL DEFINITIONS

2.8.1 *Dependency Locality Theory*

DLT (Gibson, 1998, 2000) is a theory of computational resources consumed by a human processor when parsing the syntax of a sentence. It focuses on discourse referents. It assumes that the sentence structure is parsed by the human brain one word at a time and that there is a cost of connecting a discourse referent into the built structure so far (integration cost) as well as a cost of keeping track of incomplete dependencies (memory cost). Like Demberg and Keller (2008), this thesis focus on integration cost.

The integration cost assigns one cost unit for every new discourse referent as well as one cost unit for every discourse referent between a particular discourse referent and its head. "A discourse referent is an entity that has a spatio-temporal location so that it can later be referred to with an anaphoric expression, such as a pronoun for NPs, or tense on a verb for events." (Gibson, 1998, p. 12). When automatically assigning costs on dependency parsed text, a discourse referent can be operationalised as the head of a noun phrase or of a verb phrase.

The integration cost expresses the empirically known fact that the distance between a verb and its argument has an effect on sentence processing difficulty, e.g. Gibson and Ko (1998), though there is conflicting evidence e.g. for German verb-final sentences (Konieczny, 2000). DLT describes the increased complexity when integrating *admitted* in *The reporter that the senator attacked admitted the error* compared to *The reporter admitted the error*. The linearly increasing cost must be considered an arbitrary oversimplification. Following Demberg and Keller (2008), we use the logarithm of the cost.

Part II

TWO EXPLORATORY STUDIES

3

READING BEHAVIOR PREDICTS SYNTACTIC CATEGORIES

ABSTRACT

It is well-known that readers are less likely to fixate their gaze on closed class syntactic categories such as prepositions and pronouns. This paper investigates to what extent the syntactic category of a word in context can be predicted from gaze features obtained using eye-tracking equipment. If syntax can be reliably predicted from eye movements of readers, it can speed up linguistic annotation substantially, since reading is considerably faster than doing linguistic annotation by hand. Our results show that gaze features do discriminate between most pairs of syntactic categories, and we show how we can use this to annotate words with part of speech across domains, when tag dictionaries enable us to narrow down the set of potential categories.

3.1 INTRODUCTION

Eye movements during reading is a well-established proxy for cognitive processing, and it is well-known that readers are more likely to fixate on words from open syntactic categories (verbs, nouns, adjectives) than on closed category items like prepositions and conjunctions (Nilsson and Nivre, 2009; Rayner, 1998). Generally, readers seem to be most likely to fixate and re-fixate on nouns (Furtner, Rauthmann, and Sachse, 2009). If reading behavior is affected by syntactic category, maybe reading behavior can, conversely, also tell us about the syntax of words in context.

This paper investigates to what extent gaze data can be used to *predict* syntactic categories. We show that gaze data can effectively be used to discriminate between a wide range of POS pairs, and gaze data can therefore be used to significantly improve type-constrained POS taggers. This is potentially useful, since eye-tracking data becomes more and more readily available with the emergence of eye trackers in mainstream consumer products (San Agustin et al., 2010). With the development of robust eye-tracking in laptops, it is easy to imagine digital text providers storing gaze data, which could then be used to improve automated analysis of their publications.

CONTRIBUTIONS We are, to the best of our knowledge, the first to study reading behavior of syntactically annotated, natural text across

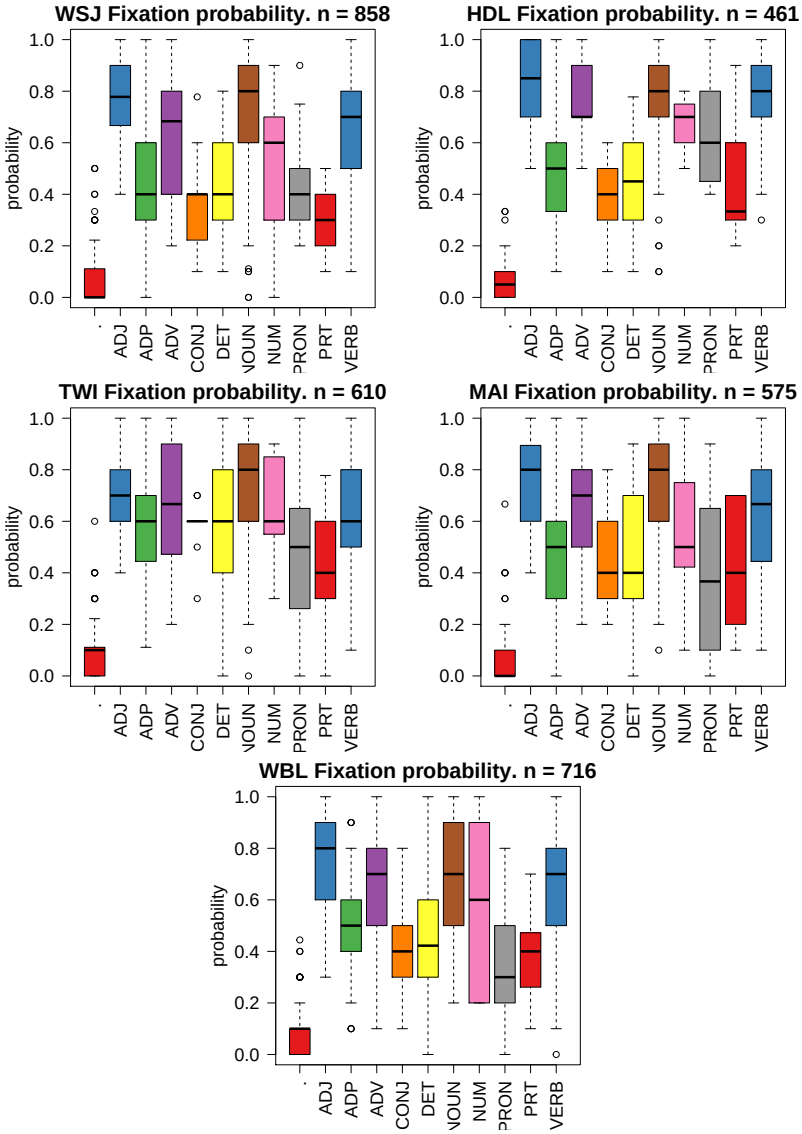


Figure 3.1: Fixation probability boxplots across five domains

domains, and how gaze correlates with a complete set of syntactic categories. We use logistic regression to show that gaze features discriminate between POS pairs, even across domains. We then show how gaze features can improve a cross-domain supervised POS tagger. We show that gaze-based predictions are robust, not only across domains, but also across subjects.

3.2 EXPERIMENT

In our experiment, 10 subjects read syntactically annotated sentences from five domains.

3.2.1 *Data*

The data consists of 250 sentences: 50 sentences (min. 3 tokens, max. 120 characters), randomly sampled from each of five different, manually annotated corpora: Wall Street Journal articles (WSJ), Wall Street Journal headlines (HDL), emails (MAI), weblogs (WBL), and Twitter (TWI). WSJ and HDL syntactically annotated sentences come from the OntoNotes 4.0 release of the English Penn Treebank.[1] The MAI and WBL sections come from the English Web Treebank.[2] The TWI data comes from the work of Foster et al. (2011). We mapped the gold labels to the 12 UPOS (Petrov, Das, and McDonald, 2011), but discarded the category X due to data sparsity.

3.2.2 *Experimental design*

The 250 items were read by all 10 participants, but participants read the items in one of five randomized orders. Neither the source domain for the sentence, nor the POS tags were revealed to the participant at any time. One sentence was presented at a time in black on a light gray background. Font face was Verdana and font size was 25 pixels. Sentences were centered vertically, and all sentences could fit into one line. All sentences were preceded by a fixation cross. The experiment was self-paced. To switch to a new sentence and to ensure that the sentence was actually processed by the participant, participants rated the immediate interest towards the sentence on a scale from 1-6 by pressing the corresponding number on the numeric keypad. Participants were instructed to read and continue to the next sentence as quickly as possible. The actual experiment was preceded by 25 practice sentences to familiarize the participant with the experimental setup.

Our apparatus was a Tobii X120 eye tracker with a 15" monitor. Sampling rate was 120 Hz binocular. Participants were seated on a chair approximately 65 cm from the display. We recruited 10 participants (7 male, mean age 31.30 $\pm$4.74)) from campus. All were native English speakers. Their vision was normal or corrected to normal, and none were diagnosed with dyslexia. All were skilled readers. Minimum educational level was an ongoing MA. Each session lasted around 40 minutes. One participant had no fixations on a few sentences. We believe that erroneous key strokes caused the participant to skip a few sentences.

1 http://catalog.ldc.upenn.edu/LDC2011T03
2 http://catalog.ldc.upenn.edu/LDC2012T13

Rank	Feature	% of votes
0	Fixation prob	19.0
1	Previous word fixated binary	13.7
2	Next word fixated binary	13.2
3	nFixations	12.2
4	First fixation duration on every word	9.1
5	Previous fixation duration	7.0
6	Mean fixation duration per word	6.6
7	Re-read prob	5.7
8	Next fixation duration	2.0
9	Total fixation duration per word	2.0

Table 3.1: 10 most used features by stability selection from logistic regression classification of all POS pairs on all domains, 5-fold cross validation.

3.2.3 *Features*

There are many different features for exploring cognitive load during reading (Rayner, 1998). We extracted a broad selection of cognitive effort features from the raw eye-tracking data in order to determine which are more fit for the task. The features are inspired by Salojärvi et al. (2003), who used a similarly exploratory approach. We wanted to cover both oculomotor features, such as fixations on previous and subsequent words, and measures relating to early (e.g. first fixation duration) and late processing (e.g. regression destinations / departure points and total fixation time). We also included reading speed and reading depth features, such as fixation probability and total fixation time per word. In total, we have 32 gaze features, where some are highly correlated (such as number of fixations on a word and total fixation time).

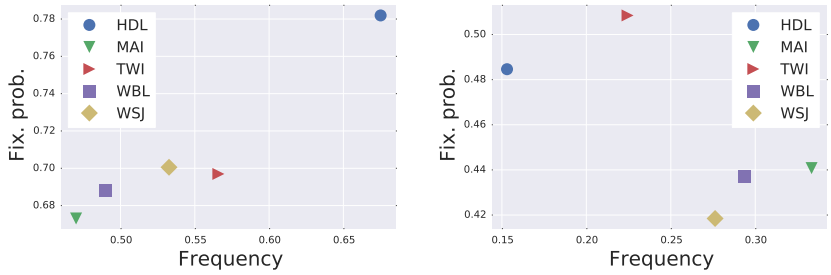
3.2.4 *Dundee Corpus*

The main weakness of the experiment is the small dataset. As future work, we plan to replicate the experiment with a $99 eye tracker for subjects to use at home. This will make it easy to collect thousands of sentences, leading to more robust gaze-based POS models. Here, instead, we include an experiment with the Dundee corpus (Kennedy and Pynte, 2005). The Dundee corpus is a widely used dataset in research on reading and consists of gaze data for 10 subjects reading 20 newswire articles (about 51,000 words). We extracted the same word-based features as above, except probability for 1st and 2nd fixation,

and sentence-level features (in the Dundee corpus, subjects are exposed to multiple sentences per screen window), and used them as features in our POS tagging experiments in Section 3.3.3.

3.2.5 *Learning experiments*

In our experiments, we used type-constrained logistic regression with L2-regularization and type-constrained (averaged) structured perceptron (Collins, 2002; Täckström et al., 2013). In all experiments, unless otherwise stated, we trained our models on four domains and evaluated on the fifth to avoid over-fitting to the characteristics of a specific domain. Our tag dictionary is from Wiktionary[3] and covers 95% of all tokens.



(a) Content words (b) Function words

Figure 3.2: Scatter plot of frequency and fixation probability for content words (NOUN, VERB, ADJ, NUM) and function words (PRON, CONJ, ADP, DET, PRT)

3.3 RESULTS

3.3.0.1 *Domain differences*

Our first observation is that the gaze characteristics differ slightly across domains, but more across POS. Figure 3.1 presents the fixation probabilities across the 11 parts of speech. While the overall pattern is similar across the five domains (open category items are more likely to be fixated), we see domain differences. For example, pronouns are more likely to be fixated in headlines. The explanation could lie in the different distributions of function words and content words. It is established and unchallenged that function words are fixated on about 35% of the time and content words are fixated on about 85% of the time (Rayner and Duffy, 1988). In our data, these numbers vary among the domains according to frequency of that word class, see Figure 3.2. Figure 3.2a shows that there is a strong linear correlation between content word frequency and content word fixation probability

3 https://code.google.com/p/wikily-supervised-pos-tagger/downloads/list

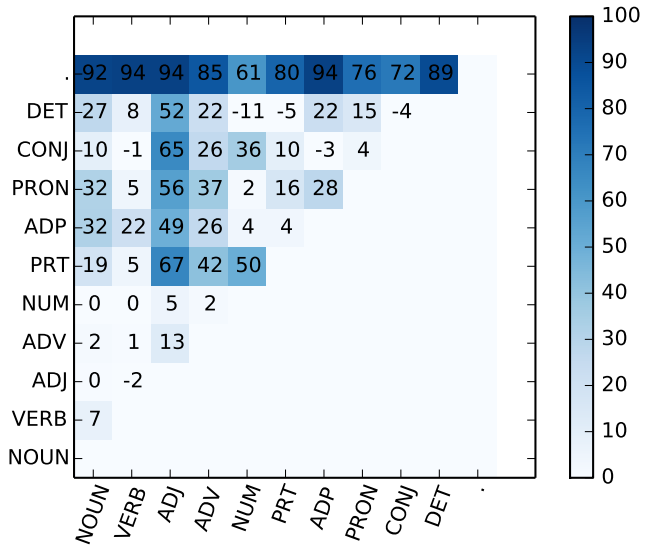


Figure 3.3: Error reduction of logistic regression over a majority baseline. All domains

among the different domains: Pearson's $\rho = 0.909$. From Figure 3.2b, there is a negative correlation between function word frequency and function word fixation probability: Pearson's $\rho = -0.702$.

3.3.1 *Predictive gaze features*

To investigate which gaze features were more predictive of part of speech, we used stability selection (Meinshausen and Bühlmann, 2010) with logistic regression classification on all binary POS classifications. Fixation probability was the most informative feature, but also whether the words around the word is fixated is important along with num-

	SP	+Gaze	+DGaze	+FreqLen	+DGaze+FreqLen
HDL	80.7	82.2	82.2	82.6	**84.3**
MAI	79.1	83.1	**83.4**	79.5	83.1
TWI	77.1	78.7	**80.0**	77.2	79.3
WBL	83.6	85.4	85.8	85.0	**86.1**
WSJ	83.1	83.7	83.8	83.1	**85.9**
Average	80.7	82.6	83.0	81.5	**83.7**

Table 3.2: POS tagging accuracy scores on different test sets using 200 out-of-domain sentences for training. SP is baseline results from the structured perceptron. DGaze is using gaze features from Dundee. Best result for each row in bold face

ber of fixations. In our binary discrimination and POS tagging experiments, using L2-regularization or averaging with all features was superior (on Twitter data) to using stability selection for feature selection. We also asked a psycholinguist to select a small set of relatively independent gaze features fit for the task (first fixation duration, fixation probability and re-read probability), but again, using all features with L2-regularization led to better performance on the Twitter data.

3.3.2 *Binary discrimination*

First, we trained L2-regularized logistic regression models to discriminate between all pairs of POS tags only using gaze features. In other words, for example we selected all words annotated as NOUN or VERB, and trained a logistic regression model to discriminate between the two in a five-fold cross validation setup. We report error reduction $\frac{acc-baseline}{1-baseline}$ in Figure 3.3.

3.3.3 *POS tagging*

We also tried evaluating our gaze features directly in a supervised POS tagger.[4] We trained a type-constrained (averaged) perceptron model with drop-out and a standard feature model (from Owoputi et al. (2013)) augmented with the above gaze features. The POS tagger was trained on a very small seed of data (200 sentences), doing 20 passes over the data, and evaluated on out-of-domain test data; training on four domains, testing on one. For the gaze features, instead of using token gaze features, we first built a lexicon with average word type statistics from the training data. We normalize the gaze matrix by dividing with its standard deviation. This is the normalizer in Turian, Ratinov, and Bengio (2010) with $\sigma = 1.0$. We condition on the gaze features of the current word, only. We compare performance using gaze features to using only word frequency, estimating from the (unlabeled) English Web Treebank corpus, and word length (FreqLen).

The first three columns in Table 3.2 show, that gaze features help POS tagging, at least when trained on very small seeds of data. Error reduction using gaze features from the Dundee corpus (DGaze) is 12%. We know that gaze features correlate with word frequency and word length, but using these features directly leads to much smaller performance gains. Concatenating the two features sets leads to the best performance, with an error reduction of 16%.

In follow-up experiments, we observe that averaging over 10 subjects when collecting gaze features does not seem as important as we expected. Tagging accuracies on raw (non-averaged) data are only about 1% lower. Finally, we also tried running logistic regression ex-

4 https://github.com/coastalcph/rungsted

periments across subjects rather than domains. Here, tagging accuracies were again comparable to our set-up, suggesting that gaze features are also robust across subjects.

3.4 RELATED WORK

Matthies and Søgaard (2013) present results that suggest that individual variation among (academically trained) subjects' reading behavior was not a greater source of error than variation within subjects, showing that it is possible to predict fixations across readers. Our work relates to such work, studying the robustness of reading models across domains and readers, but it also relates in spirit to research on using weak supervision in NLP, e.g., work on using HTML markup to improve dependency parsers (Spitkovsky, 2013) or using click-through data to improve POS taggers (Ganchev et al., 2012).

3.5 CONCLUSIONS

We have shown that it is possible to use gaze features to discriminate between many POS pairs across domains, even with only a small dataset and a small set of subjects. We also showed that gaze features can improve the performance of a POS tagger trained on small seeds of data.

4

USING READING BEHAVIOR TO PREDICT GRAMMATICAL FUNCTIONS

ABSTRACT

This paper investigates to what extent grammatical functions of a word can be predicted from gaze features obtained using eye-tracking. A recent study showed that reading behavior can be used to predict coarse-grained part of speech, but we go beyond this, and show that gaze features can also be used to make more fine-grained distinctions between grammatical functions, e.g., subjects and objects. In addition, we show that gaze features can be used to improve a discriminative transition-based dependency parser.

4.1 INTRODUCTION

Readers fixate more and longer on open syntactic categories (verbs, nouns, adjectives) than on closed class items like prepositions and conjunctions (Nilsson and Nivre, 2009; Rayner and Duffy, 1988). Recently, Barrett and Søgaard (2015a) presented evidence that gaze features can be used to discriminate between most pairs of POS. Their study uses all the coarse-grained POS labels proposed by Petrov, Das, and McDonald (2011). This paper investigates to what extent gaze data can also be used to predict grammatical functions such as subjects and objects. We first show that a simple logistic regression classifier trained on a very small seed of data using gaze features discriminates between some pairs of grammatical functions. We show that the same kind of classifier distinguishes well between the four main grammatical functions of nouns, POBJ, DOBJ, NN, and NSUBJ. In

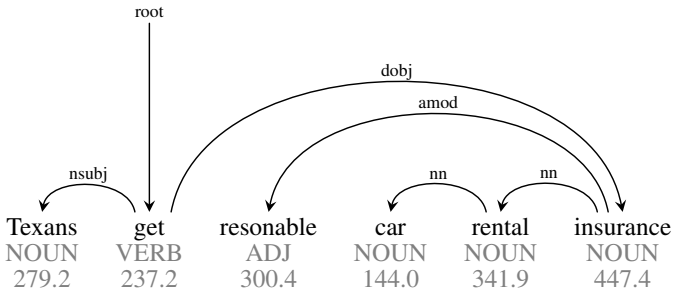


Figure 4.1: A dependency structure with average fixation duration per word.

Section 4.4.4, we also show how gaze features can be used to improve dependency parsing. Many gaze features correlate with word length and word frequency (Rayner, 1998) and these could be as good as gaze features, while being easier to obtain. We use frequencies from the unlabeled portions of the English Web Treebank and word length as baseline in all types of experiments and find that gaze features to be better predictors for the noun experiment as well as for improving parsers.

This work is of psycholinguistic interest, but we show that gaze features may have practical relevance, by demonstrating that they can be used to improve a dependency parser. Eye-tracking data becomes more readily available with the emergence of eye trackers in mainstream consumer products (San Agustin et al., 2010). With the development of robust eye-tracking in laptops, it is easy to imagine digital text providers storing gaze data, which could then be used as partial annotation of their publications.

CONTRIBUTIONS We demonstrate that we can discriminate between some grammatical functions using gaze features and which features are fit for the task. We show a practical use for data reflecting human cognitive processing. Finally, we use gaze features to improve a transition-based dependency parser, comparing also to dependency parsers augmented with word embeddings.

4.2 EYE TRACKING DATA

The data comes from Barrett and Søgaard (2015a) and is publicly available[1]. In this experiment 10 native English speakers read 250 syntactically annotated sentences in English (min. 3 tokens, max. 120 characters). The sentences were randomly sampled from one of five different, manually annotated corpora from different domains: Wall Street Journal articles (WSJ), Wall Street Journal headlines (HDL), emails (MAI), weblogs (WBL), and Twitter (TWI)[2]. See Figure 4.1 for an example.

FEATURES It is not yet established which eye movement reading features are fit for the task of distinguishing grammatical functions of the words. To explore this, we extracted a broad selection of word- and sentence-based features. The features are inspired by Salojärvi et al. (2003) who used a similar exploratory approach. For a full list of features, see Appendix A.

1 https://bitbucket.org/lowlands/release/src

2 Wall Street Journal sentences are from OntoNotes 4.0 release of the English Penn Treebank. http://catalog.ldc.upenn.edu/LDC2011T03. Mail and weblog sentences come from the English Web Treebank. http://catalog.ldc.upenn.edu/LDC2012T13. Twitter sentences are from the work of (Foster et al., 2011)

4.3 LEARNING EXPERIMENTS

In our binary experiments, we use L2-regularized logistic regression classifiers with the default parameter setting in SciKit Learn[3] and a publicly available transition-based dependency parser[4] trained using structured perceptron (Collins, 2002; Zhang and Nivre, 2011).

4.3.1 *Binary classification*

We trained logistic regression models to discriminate between pairs of the 11 most frequent dependency relations where the sample size is above 100: (AMOD, NN, AUX, PREP, NSUBJ, ADVMOD, DEP, DET, DOBJ, POBJ, ROOT) only using gaze features. E.g., we selected all words annotated as PREP or NSUBJ and trained a logistic regression model to discriminate between the two in a five-fold cross validation setup. Our baseline uses the following features: word length, position in sentence and word frequency.

Some dependency relations are almost uniquely associated with one POS, e.g. determiners where 84.8% of words with the dependency relation DET are labeled determiners. This means that in some cases, the grammatical function of a word follows from its part of speech. In another binary experiment, we therefore focus on nouns to show that eye movements *do* make more fine-grained distinctions between different grammatical functions. Nouns are mostly four-way ambiguous: 74.6% of the 946 nouns in the dataset have one of four dependency relations to its head. Nouns with POBJ relations is 18.9% of all nouns, NSUBJ is 17.0%, NN is 27.0% and DOBJ is 14.9%. The remaining 25.4% of the nouns are discarded from the noun experiment since they have 28 different relations to their head.

4.3.2 *Parsing*

In all experiments we trained our parsing models on four domains and evaluated on the fifth to avoid over-fitting to the characteristics of a specific domain. All parameters were tuned on the WSJ dataset. We did 30 passes over the data and used the feature model in Zhang and Nivre (2011) – concatenated with gaze vectors for the first token on the buffer, the first token in the stack, and the left sibling of the first token in the stack. We extend the feature representation of each parser configuration by 3×26 features. Our gaze vectors were normalized using the technique in Turian, Ratinov, and Bengio (2010) ($\sigma \cdot E/SD(E)$) using a scaling factor of $\sigma = 0.001$. Gaze features such as fixation duration are known to correlate with word frequency and

3 http://scikit-learn.org/stable/modules/generated/sklearn.linear_model.LogisticRegression.html

4 https://github.com/andersjo/hanstholm

Rank	Feature name	% of votes
0	Next word fixation probability	13.46
1	Fixation probability	11.14
2	*n* Fixations	9.66
3	Probability to get 2nd fixation	8.90
4	Previous word fixation probability	7.17
5	*n* Regressions from	5.65
6	First fixation duration on every word	5.45
7	Mean fixation duration per word	5.17
8	Previous fixation duration	4.93
9	Re-read probability	4.65
10	Probability to get 1st fixation	4.53
11	*n* Long regressions from word	3.77
12	Share of fixated words per sent	3.04
13	*n* Re-fixations	1.88
14	*n* Regressions to word	1.76

Table 4.1: Most predictive features for binary classification of 11 most frequent dependency relations using five-fold cross validation.

word length. To investigate whether word length and frequency are stronger features than gaze, we perform an experiment, +Freq+Len, where our baseline and system also use frequencies and word length as features.

4.4 Results

4.4.1 *Predictive features*

To investigate which gaze features were more predictive of grammatical function, we used stability selection (Meinshausen and Bühlmann, 2010) with logistic regression classification on binary dependency relation classifications on the most frequent dependency relations.

For each pair of dependencies, we perform a five-fold cross validation and record the informative features from each run. 4.1 shows the 15 most used features in ranked order with their proportion of all votes. The features predictive of grammatical functions are similar to the features that were found to be predictive of POS (Barrett and Søgaard, 2015a), however, the probability that a word gets first and second fixation were not important features for POS classification, whereas they are contributing to dependency classification. This could suggest that words with certain grammatical functions are con-

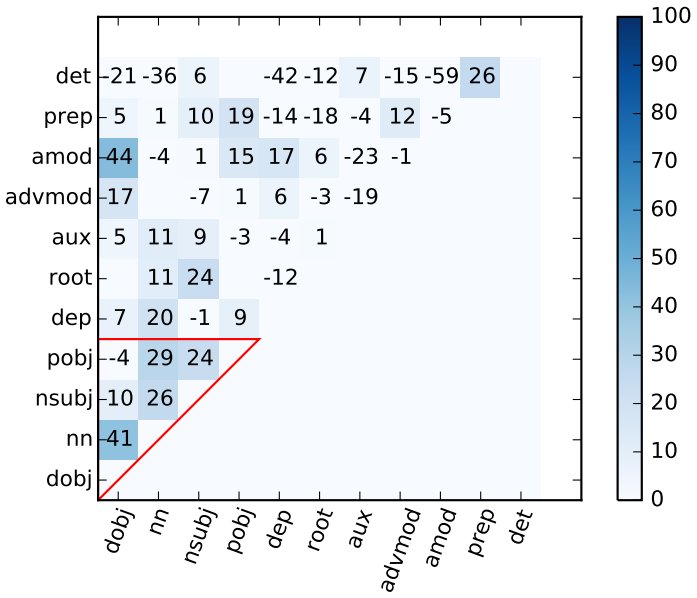


Figure 4.2: Error reduction over the baseline for binary classifications of 11 most frequent dependency relations. 5-fold cross validation. Dependency relations associated with nouns in triangle.

sistently more likely or less likely to get first and second fixation, but could also be due to a frequent syntactic order in the sample.

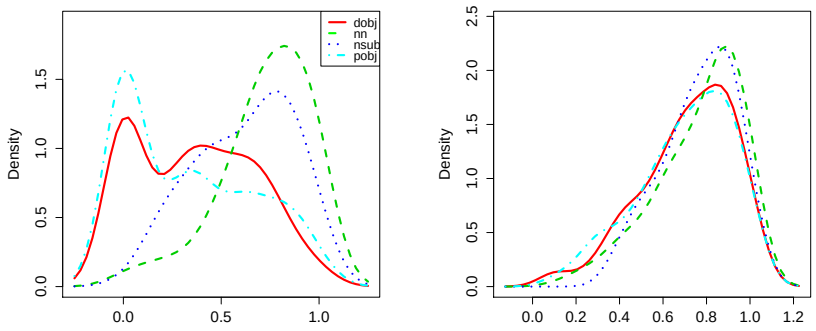
4.4.2 *Binary discrimination*

Error reduction over the baseline can be seen in Figure 4.2. The mean accuracy using logistic regression on all binary classification problems between grammatical functions is 0.722. The frequency-position-word length baseline is 0.706. In other words, using gaze features leads to a 5.6% error reduction over the baseline. The worst performance (where our baseline outperforms using gaze features) is seen where one relation is associated with closed class words (DET, PREP, AUX), and where discrimination is easier.

4.4.3 *Noun experiment*

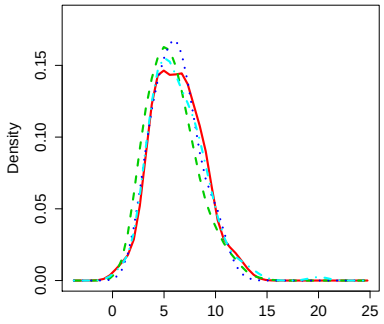
Error reductions for pairwise classification of nouns are between -4% and 41%. See Figure 4.2. The average accuracy for binary noun experiments is 0.721. Baseline accuracy is 0.647. For POBJ and DOBJ the baseline was better than using gaze, but for the other pairs, gaze was better. When doing stability selection for nouns with only the four most frequent grammatical functions, the most important features can be seen from Table 4.2. The most informative feature is the fixation probability of the next word. Kernel density of this feature can be seen in Figure 4.3a, and it shows two types of behavior: POBJ

and DOBJ, where the next word is less frequently fixated, and NN and NSUBJ, where the next word is more frequently fixated. Whether the next word is fixated or not, can be influenced by the word length, as well as the fixation probability of the current word: If the word is very short, the next word can be processed from a fixation of the current word, and if the current word is not fixated, the eyes need to land somewhere in order for the visual span to cover a satisfactory part of the text. Word length and fixation probabilities for the nouns are reported in Figure 4.3c and Figure 4.3b to show that the dependency labels have similar densities.



(a) Next word fixation probability

(b) Fixation probability



(c) Word length

Figure 4.3: Kernel density plots across four grammatical functions of nouns.

4.4.4 *Dependency parsing*

We also evaluate our gaze features directly in a supervised dependency parser. Our baseline performance is relatively low because of the small training set, but comparable to performance often seen with low-resource languages. Evaluation metrics are labelled attachment score (LAS) and unlabelled attachment score (UAS), i.e. the number of

RANK	FEATURE NAME	% OF VOTES
0	Next word fixation probability	20.66
1	Probability to get 2nd fixation	19.83
2	nRegressions from word	14.05
3	Previous word fixation probability	8.68
4	Probability to get 1st fixation	7.44

Table 4.2: Most predictive features for the binary classification of four most frequent dependency relations for nouns using five-fold cross validation.

	LAS						UAS					
					+FREQ+LEN						FREQ+LEN	
	BL	+SENNA	+EIGENW	+GAZE	BL	+GAZE	BL	+SENNA	+EIGENW	+GAZE	BL	+GAZE
HDL	53.9	53.9	52.6	**54.1**	53.5	**54.2**	58.3	**60.0**	56.4	58.9	58.2	**58.7**
MAI	66.7	65.1	66.8	***68.4**	67.8	***71.1**	71.5	69.9	71.5	***74.7**	73.2	***75.9**
TWI	53.2	56.9	**56.3**	*56.1	55.4	***56.9**	57.6	**62.6**	61.5	*60.2	60.7	***62.1**
WBL	60.4	62.9	59.2	***63.8**	63.1	***65.5**	66.8	67.0	66.6	***71.1**	70.9	***71.9**
WSJ	63.5	63.5	62.2	***65.0**	62.9	**63.4**	67.2	68.1	67.4	***69.5**	67.1	**67.7**
Average	59.5	60.5	59.4	***61.5**	60.5	***62.2**	64.3	65.5	64.7	***66.9**	66.0	***67.2**

Table 4.3: Dependency parsing results on all five test sets using 200 sentences (four domains) for training and 50 sentences (one domain) for evaluation. Best results are bold-faced, and significant ($p < 0.01$) improvements are asterisked.

words that get assigned the correct syntactic head w/o the correct dependency label.

Gaze features lead to consistent improvements across all five domains. The average error reduction in LAS is 5.0%, while the average error reduction in UAS is 7.3%. For the +FREQ+LEN experiment, +GAZE also lead to improvements for all domains, with error reductions of 3.3% for LAS and 4.7% for UAS.

For comparison we also ran our parser with SENNA embeddings[5] and EIGENWORDS embeddings.[6] The gaze vectors proved overall more informative.

4.5 RELATED WORK

In addition to Barrett and Søgaard (2015a), our work relates to Matthies and Søgaard (2013), who study the robustness of a fixation prediction model across readers, not domains, but our work also relates in spirit to research on using weak supervision in NLP, e.g., work on using HTML markup to improve dependency parsers (Spitkovsky, 2013)

5 http://ronan.collobert.com/senna/
6 http://www.cis.upenn.edu/~ungar/eigenwords/

or using click-through data to improve POS taggers (Ganchev et al., 2012).

There have been few studies correlating reading behavior and general dependency syntax in the literature. Demberg and Keller (2008), having parsed the Dundee corpus using MiniPar, show that dependency integration cost, roughly the distance between a word and its head, is predictive of reading times for nouns. Our finding could be a side-effect of this, since NSUBJ, NN and DOBJ/POBJ typically have very different dependency integration costs, while DOBJ and POBJ have about the same. Their study thus seems to support our finding that gaze features can be used to discriminate between the grammatical functions of nouns. Most other work of this kind focus on specific phenomena, e.g., Traxler, Morris, and Seely (2002), who show that subjects find it harder to process object relative clauses than subject relative clauses. This paper is related to such work, but our interest is a broader model of syntactic influences on reading patterns.

4.6 CONCLUSIONS

We have shown that gaze features can be used to discriminate between a subset of grammatical functions, even across domains, using only a small dataset and explored which features are more useful. Furthermore, we have shown that gaze features can be used to improve a state-of-the-art dependency parsing model, even when trained on small seeds of data, which suggests that parsers can benefit from data from human processing.

Part III

PART-OF-SPEECH INDUCTION USING GAZE AND OTHER HUMAN TEXT PROCESSING DATA

5

THE DUNDEE TREEBANK

ABSTRACT

We introduce the Dundee Treebank, a Universal Dependencies-style syntactic annotation layer on top of the English side of the Dundee Corpus. As the Dundee Corpus is an important resource for conducting large-scale psycholinguistic research, we aim at facilitating further research in the field by replacing automatic parses with manually assigned syntax. We report on constructing the treebank, performing parsing experiments, as well as replicating a broad-scale psycholinguistic study – now for the first time using manually assigned syntactic dependencies.

5.1 INTRODUCTION

The Dundee Corpus is a major resource for studies of linguistic processing through eye movements. It is a famous resource in psycholinguistics, and – to the best of our knowledge – the world's largest eye-movement corpus. The English part of the Dundee Corpus was annotated with POS information in 2009 (Frank, 2009). This layer of annotation facilitated new psycholinguistic studies such as testing several reader models using models of hierarchical phrase structure and sequential structure (Frank and Bod, 2011).

In this paper, we describe a recent annotation effort to add a layer of dependency syntax on top of the POS annotation, enabling the replication of classic studies such as (Demberg and Keller, 2008) on manually assigned syntax rather than automatic parses. We first describe the Dundee Corpus, then our annotation scheme, and finally we discuss applications of this annotation effort.

5.2 THE DUNDEE CORPUS

The Dundee Corpus was developed by Alan Kennedy and Joël Pynte in 2003, and it contains eye movement data on top of English and French text (Kennedy, Hill, and Pynte, 2003). Measurements were taken while participants read newspaper articles from *The Independent* (English) or *Le Monde* (French). Ten native English-speaking subjects participated in the English experiments reading 20 articles, which we focus on here. For a more detailed account, see Kennedy and Pynte (2005).

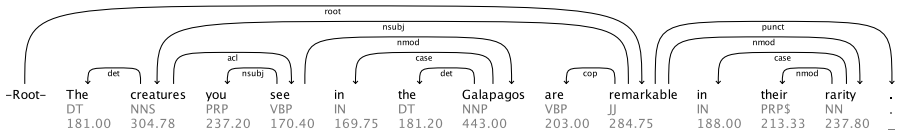


Figure 5.1: An example sentence (#10) from the Dundee Corpus with UD-style syntactic dependencies and per-word fixation durations.

The English corpus contains 51,502 tokens[1] and 9,776 types in 2,368 sentences. The apparatus was a Dr Bouis Oculometer Eye tracker with a 1000 Hz monocular (right) sampling. The corpus provides information on fixation durations and fixation order on word level – while also accounting for landing position – for a relatively natural reading scenario. Subjects read running text, 5 lines per display.

Eye movements provide a window to the workings of the brain, e.g. by reflecting cognitive load. Recordings of eye movements during reading is one of the main methods for getting a millisecond to millisecond record of human cognition. Eye movements during reading is controlled by a complex interplay between low-level factors (how much the eye can see and encode from each fixation, word length, landing position, etc.) and high-level factors (e.g. syntactic processing). For an overview, see Rayner (1998).

This resource has enabled researchers to study things like syntactic and semantic factors in processing difficulty of words (Mitchell et al., 2010) and whether the linguistic processing associated with a word can proceed before the word is uniquely identified (Smith and Levy, 2010).

5.3 SYNTACTIC ANNOTATION

In annotating the Dundee Corpus for syntactic dependencies, we follow the Universal Dependencies (UD) guidelines[2] (Agić et al., 2015) as the emerging *de facto* standard for dependency annotation.

The guidelines build on – and closely adhere to – Universal Stanford dependencies (de Marneffe et al., 2014), proposing 40 dependency relations together with an UPOS tagset and morphological features. We convert the Penn Treebank-style POS tags from the Dundee Corpus into UPOS, and we provide the universal morphology features, by using the official English UD conversion tools.

The guidelines for annotating English are very well-documented within the UD framework. We only briefly touch upon the most important ones.

For core dependents of clausal predicates, UD distinguishes between nominal subjects (NSUBJ), direct objects (DOBJ), indirect objects

1 According to the tokenisation of the Dundee corpus where punctuation and contracted words are glued to the preceding word.

2 `http://universaldependencies.github.io/`

	Dundee			English UD dev			English UD test		
Train set	LAS	UAS	LA	LAS	UAS	LA	LAS	UAS	LA
Dundee	82.23	85.06	89.97	69.50	75.96	81.26	68.86	75.60	80.61
English UD	71.45	78.66	84.28	85.51	88.03	92.91	84.72	87.30	92.37

Table 5.1: Dependency parsing results with English UD and Dundee as training sets. Parser: `mate-tools` graph-based parser with default settings (Bohnet, 2010). Features: FORM and coarse part-of-speech (CPOS) TAG only, using the Penn Treebank POS tags. Metrics: labeled and unlabeled attachment scores (LAS, UAS), and label assignment (LA). On Dundee, result is on 5-fold (80:20) cross-validation, as the Dundee Treebank has no held-out test set.

(IOBJ), nominal subjects of passives (NSUBJPASS), clausal subjects (CSUBJ), clausal subjects of passives (CSUBJPASS), clausal complements (CCOMP), and open clausal complements (XCOMP). When it comes to non-nominal modifiers of nouns, for example, the guidelines distinguishes between adjectival modifiers (AMOD), determiners (DET), and negation (NEG).

We show an example sentence from the treebank in Figure 5.1. It depicts the UD-style dependency annotation, as well as per-word total fixation durations averaged over ten readers. Some of the typical UD-style conventions – such as content head primacy and no copula heads – are also illustrated.

We used two professional annotators that had previously worked on treebanks following the UD guidelines. The annotators provided double annotations for 118 sentences, with moderately high inter-annotator agreements of 80.82 (LAS), 87.61 (UAS, and 86.63 (LA). The remaining part of the Dundee Corpus was only annotated by one annotator.

Further, we trained a graph-based dependency parser (Bohnet, 2010) on English UD training data, and parsed the Dundee Corpus text. We report the results in Table 5.1. There is a decrease in accuracy moving from English UD to the Dundee Corpus text. We attribute the decrease to the domain shift – English UD stemming from various web sources, while Dundee consists of newswire commentaries in specific – and possibly to the slight cross-dataset inconsistency in POS and dependency annotations. In a separate experiment, we also parse the Dundee Corpus text using 5-fold cross-validation with an 80:20 split, observing accuracies consistent with the English UD experiment. These results are also reported in Table 5.1.

The cross-dataset decrease in parsing accuracy, even if irrelevant for Dundee-specific experiments, plays into the argument for using gold-standard annotations in psycholinguistic research.

5.4 REPLICATION OF DEPENDENCY LOCALITY THEORY EXPERIMENT

The Dundee Treebank annotated with dependencies has the following affordances. First, it allows for replication of studies such as Demberg and Keller (2008) with manual annotations. Second, gaze features can be used to improve NLP models by enabling joint learning of gaze and syntactic dependencies (Barrett and Søgaard, 2015a,b). Finally, the Dundee Treebank facilitates for researchers to study the reading of very specific syntactic constructions in naturalistic, contextualized text, while controlling for individual variation, and variation specific to the parts of speech or syntactic dependencies involved.

Demberg and Keller (2008) were the first to test broad-covering theories of sentence processing on large-scale, contextualized text with eye tracking data (Demberg and Keller, 2008). They explored two theories of syntactic complexity, namely dependency locality theory (DLT) and Surprisal, and how these correlate with three eye tracking measures while controlling for oculomotor and low-level processing.

DLT (Gibson, 2000) estimates the computational resources consumed by the human processor and computes a cost for any discourse referent as well as a cost for every discourse referent between a particular discourse referent and it's head. Thus, DLT needs dependency parsed text to score the complexity of the sentences and Minipar was used to parse the text with a reported 83% accuracy of the DLT score.

In this paper we replicate the parts of their experiments involving DLT, but with manually assigned dependencies instead of automatic parses for calculating DLT. Demberg and Keller (2008) found that DLT score did not have the expected positive effect on reading time of all words. The calculation of DLT only applies for nouns and verbs. They did, however, find that DLT significantly had a positive effect on reading times for nouns and verbs.

We replicate the linear mixed-effects experiment using first pass fixation duration per word for all words and nouns[3]. First pass fixation duration is the duration of all fixations on a specific word from when the reader's eyes first enter into the region and until the eyes leave the region, given that this region is fixated. This is a measure said to encompass early syntactic and semantic processing as well as lexical access. We use the same low-level predictor variables as the original experiment:

1. word length in characters (WORD LENGTH),
2. log-transformed frequency of target word (WORD FREQUENCY),
3. log-transformed frequency of previous word (PREVIOUS WORD FREQUENCY),

3 The original paper does not contain information about the elements of the model for verbs, which is why this part of the experiment was not replicated.

Predictor	Coef	p	Coef original	p original
INTERCEPT	199.59		128.24	***
WORD LENGTH	-1.25		30.90	***
WORD FREQUENCY	4.43	***	14.50	***
PREVIOUS WORD FIXATED	-33.32	***	-18.05	***
LANDING POSITION	-1.23	***	-4.18	***
LAUNCH DISTANCE	1.79	***	-1.91	***
SENTENCE POSITION	-.09	*	-.12	*
FORWARD TRANSITIONAL PROBABILITY	1.51	***	-3.27	***
BACKWARD TRANSITIONAL PROBABILITY	-5.87	***	3.96	***
log(DLT)	3.51	**	5.86	*
WORD LENGTH:WORD FREQUENCY	-2.96	***	-4.98	***
WORD LENGTH:LANDING POSITION	-.68	***	-1.02	***

Table 5.2: First pass durations for nouns with non-zero DLT score in the Dundee corpus. Coefficients and their significance level. Same predictors as original noun experiment. * $p < .05$, ** $p < .01$, *** $p < .001$.

4. forward-transitional probability (FORWARD TRANSITIONAL PROBABILITY),
5. backward transitional probability (BACKWARD TRANSITIONAL PROBABILITY),
6. word position in sentence (SENTENCEPOSITION),
7. whether the previous word was fixated or not (PREVIOUS WORD FIXATED),
8. launch distance of the fixation in characters (LAUNCH DISTANCE),
9. and fixation landing position (LANDING POSITION).

Backward- and forward transitional probabilities are conditional probabilities of a word given the previous / next word, respectively (McDonald and Shillcock, 2003). Along with the word frequencies these two measures are obtained from the British National Corpus (BNC) (Consortium et al., 2007), following the line of Demberg and Keller (2008). We use KenLM (Heafield, 2011) for getting the bigram frequencies and Kneser-Ney smoothing for those bigrams that are not found in the training set. Demberg and Keller (2008) used CMU-Cambridge Language Modeling Toolkit and applied Witten-Bell smoothing. Bigrams respect sentence boundaries.

We clean the data following the described approach by using only fixated words, excluding words that are followed by any kind of punctuation and excluding first and last words of each line. We did, however, not remove words "in a region of 4 or more adjacent words that had not been fixated", since it is unclear what a "region" is (non-fixated words are already removed). This left us with 209,010 data points. Demberg and Keller (2008) report to have 200,684 data points

after cleaning. The difference is probably accounted for by the missing, last cleaning step.

We use R (R Core Team, 2015) and lme4 (Bates et al., 2015) to fit a linear mixed-effects model. In the following we use the same fixed and random effects as their models minimised using Akaike Information Criterion (AIC). The authors do not report which significance test they used. We use likelihood ratio tests of the full model with the particular fixed effect against the model without the particular fixed effect.

Demberg and Keller (2008) find that for all words, DLT had a significant, negative effect on first pass fixation duration ($p < .001$), which is a displeasing counter-intuitive result. It means higher DLT score gives a shorter fixation duration. We also find a very small negative effect (-.03) of DLT on first pass fixation duration for all words, but it doesn't reach significance. Following the original experiment, we fit a model for the nouns with non-zero DLT score, encompassing 51,786 data points. The original experiment report having 45,038. In Table 5.2 we report the coefficients and significance level for all fixed effects of this model as well as the corresponding results of the original experiment. Like the original experiment, we find that the log(DLT) had a significant positive effect on reading time ($p < .01$). These two experiments demonstrate that parser bias did not skew the results substantially.

5.5 CONCLUSION

We introduced the Dundee Treebank – a new resource for corpus-based psycholinguistic experiments. The treebank is annotated in compliance with the Universal Dependencies scheme. We presented the design choices together with a batch of dependency parsing experiments.

We also partly replicated a study, which explores how a theory of sentence complexity, DLT, is reflected in reading times. We used manually assigned dependencies instead of parsed dependencies. Like the original experiment, we found both a small negative effect of DLT on all word and a significant positive effect of DLT on reading time for nouns with non-zero DLT score.

The treebank is made publicly available for research purposes.[4]

5.6 ERRATA

During the write-up process of the thesis, I was made aware that there was a systematic error in my calculation of DLT. Thanks for Scarlett Hao for pointing that out. Unfortunately my old cleaning script is missing. Trying to replicate the cleaning I get around 218,000 data points instead of the around 209.000 data points we used last. Going

4 https://bitbucket.org/lowlands/release

Predictor	Coef	p	Coef	p
INTERCEPT	283.70		128.24	***
WORD LENGTH	-16.91	***	30.90	***
WORD FREQUENCY	16.93	***	14.50	***
PREVIOUS WORD FIXATED	-40.11	***	-18.05	***
LANDING POSITION	-.64		-4.18	***
LAUNCH DISTANCE	1.73	***	-1.91	***
SENTENCE POSITION	-.14	***	-.12	*
FORWARD TRANSITIONAL PROBABILITY	-.96	***	-3.27	***
BACKWARD TRANSITIONAL PROBABILITY	-3.80	***	3.96	***
log(DLT)	14.23	***	5.86	*
WORDLENGTH:WORDFREQUENCY	-5.62	***	-4.98	***
WORDLENGTH:LANDINGPOSITION	-.41	***	-1.02	***

Table 5.3: First pass durations for nouns with non-zero DLT score in the Dundee corpus. Corrected numbers. Coefficients and their significance level. Corrected scores. Same predictors as original noun experiment. * $p < .05$, ** $p < .01$, *** $p < .001$.

over my old code, I found a bug in the Dundee feature extraction: for around 5% of the words that were refixated in the first pass reading, re-fixation durations were added twice instead of once, making the value too high for these words.

I therefore replicate the linear mixed effects models with the corrected DLT score and the corrected First Pass Duration. For all words there is now a small, positive effect (5.82) of log(DLT) on First pass duration. The effect is significant ($p > 0.001$). Both Demberg and Keller (2008) and the original version of this study found a very small, negative effect of DLT, which in our case was not significant. The correction follow the same tendency further. Though the effect is very small, it is a pleasing conclusion to have a positive effect of DLT. Due to the different cleaning procedure, the result is not directly comparable to previous studies.

For nouns only there is now a larger positive effect on the slope (14.23) of DLT than we found previously (3.51). It is also larger than what Demberg and Keller (2008) found (5.86). Table 5.3 presents the corrected coefficients.

Shain et al. (2016a) also found a small, negative effect of several variations of DLT on all words using their own later annotation of the Dundee Corpus, but it is not clear from the abstract how they cleaned the data, which statistical test they used or which fixed effect they used, making comparison difficult.

6

WEAKLY SUPERVISED PART-OF-SPEECH TAGGING USING EYE-TRACKING DATA

ABSTRACT

For many of the world's languages, there are no or very few linguistically annotated resources. On the other hand, raw text, and often also dictionaries, can be harvested from the web for many of these languages, and part-of-speech taggers can be trained with these resources. At the same time, previous research shows that eye-tracking data, which can be obtained without explicit annotation, contains clues to part-of-speech information. In this work, we bring these two ideas together and show that given raw text, a dictionary, and eye-tracking data obtained from naive participants reading text, we can train a weakly supervised part-of-speech tagger using a second-order hidden Markov model with maximum entropy emissions (SHMM-ME). The best model use type-level aggregates of eye-tracking data and significantly outperforms a baseline that does not have access to eye-tracking data.

6.1 INTRODUCTION

According to Ethnologue, there are around 7,000 languages in the world.[1] For most of these languages, no or very little linguistically annotated resources are available. This is why over the past decade or so, NLP researchers have focused on developing unsupervised algorithms that learn from raw text, which for many languages is widely available on the web. An example is part-of-speech (POS) tagging, in which unsupervised approaches have been increasingly successful (see Christodoulopoulos, Goldwater, and Steedman (2010) for an overview). The performance of unsupervised POS taggers can be improved further if dictionary information is available, making it possible to constrain the POS tagging process. Again, dictionary information can be harvested readily from the web for many languages (Li, Graça, and Taskar, 2012).

In this paper, we show that POS tagging performance can be improved further by using a weakly supervised model which exploits eye-tracking data in addition to raw text and dictionary information. Eye-tracking data can be obtained by getting native speakers of the target language to read text while their gaze behavior is recorded. Reading is substantially faster than manual annotation, and compe-

1 http://www.ethnologue.com/world

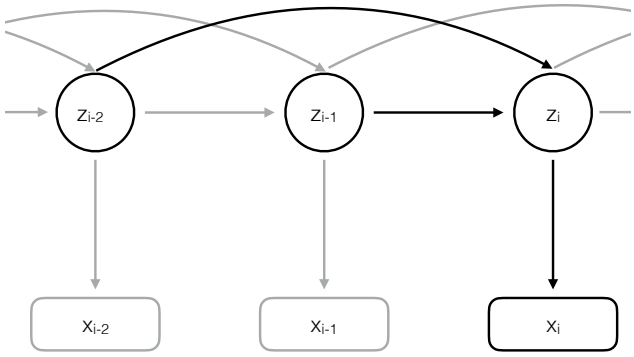


Figure 6.1: Second-order HMM. In addition to the transitional probabilities of the antecedent state z_{i-1} in first-order HMMs, second-order models incorporate transitional probabilities from the second-order antecedent state z_{i-2}.

tent readers are available for languages where trained annotators are hard to find or non-existent. While high quality eye-tracking equipment is still expensive, $100 eye-trackers such as the EyeTribe are already on the market, and cheap eye-tracking equipment is likely to be widely available in the near future, including eye-tracking by smartphone or webcam (Skovsgaard, Hansen, and Møllenbach, 2013; Xu et al., 2015).

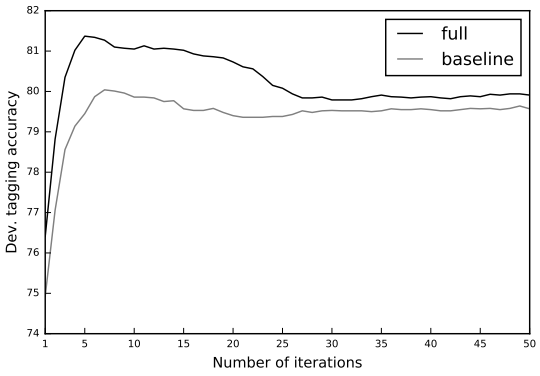


Figure 6.2: Tagging accuracy on development data (token-level) as a function of number of iterations on baseline and full model.

Gaze patterns during reading are strongly influenced by the parts of speech of the words being read. Psycholinguistic experiments show that readers are less likely to fixate on closed-class words that are predictable from context. Readers also fixate longer on rare words, on words that are semantically ambiguous, and on words that are morphologically complex (Rayner, 1998). These findings indicate that eye-tracking data should be useful for classifying words by part of

EARLY	First fixation duration
	w-1 fixation probability
	w-1 fixation duration
	First pass duration
LATE	Total regression-to duration
	n long regressions to w
	n refixations
	Re-read probability
	n regressions to w
BASIC	Total fixation duration
	Mean fixation duration
	n fixations
	Fixation probability
REGFR.	n regressions from w
	n long regressions from w
	Total regression-from duration
CONTEXT	w+1 fixation probability
	w+1 fixation duration
	w+2 fixation probability
	w+2 fixation duration
	w-2 fixation probability
	w-2 fixation probability
NOGAZEB.	Word length
	BNC log frequency
	w-1 BNC log frequency
	BNC forward transitional log probability
	BNC backward transitional log probability
NOGAZED.	Word length
	Dundee log frequency
	w-1 Dundee log frequency
	Dundee forward transitional log probability
	Dundee backward transitional log probability

Table 6.1: Features in feature selection groups.

speech, and indeed Barrett and Søgaard (2015a) show that word-type-level aggregate statistics collected from eye-tracking corpora can be used as features for supervised POS tagging, leading to substantial gains in accuracy across domains. This leads us to hypothesize that gaze data should also improve weakly supervised POS tagging.

In this paper, we test this hypothesis by experimenting with a POS tagging model that uses raw text, dictionary information, and eye-tracking data, but requires no explicit annotation. We start with a

FEATURES	TA
NOGAZEDUN	81.03
NOGAZEBNC	80.69
BASIC	80.30
EARLY	79.96
LATE	79.87
REGFROM	79.62
CONTEXT	79.53
Best Group Comb (All)	81.37
Best Gaze-Only Comb (BASIC-LATE)	80.45

Table 6.2: Tagging accuracy on the development set (token-level) for all individual feature groups, for the best combination of groups and for the best gaze-only combination of groups.

state-of-the-art unsupervised POS tagging model, the SHMM-ME of Li, Graça, and Taskar (2012), which uses only textual features. We augment this model with a wide range of features derived from an eye-tracking corpus at training time (type-level gaze features). We also experiment with token-level gaze features; the use of these features implies that eye-tracking is available both at training time and at test time. We find that eye-tracking features lead to a significant increase in POS tagging accuracy, and that type-level aggregates work better than token-level features.

6.2 THE DUNDEE TREEBANK

The Dundee Treebank (Barrett, Agić, and Søgaard, 2015) is a Universal Dependencies (UD) annotation layer that has recently been added to the world's largest eye-tracking corpus, called the Dundee Corpus (Kennedy, Hill, and Pynte, 2003). The English portion of the corpus contains 51,502 tokens and 9,776 types in 2,368 sentences. The Dundee Corpus is a well-known and widely used resource in psycholinguistic research. The corpus enables researchers to study the reading of contextualized, running text obtained under relatively naturalistic conditions. The eye-movements in the Dundee Corpus were recorded with a high-end eye-tracker, sampling at 1000 Hz. The corpus contains the eye-movements of ten native English speakers as they read the same twenty newspaper articles from *The Independent*. The corpus was augmented with Penn Treebank POS annotation by Frank (2009). When constructing the Dundee Treebank, this POS annotation was checked and corrected if necessary. In the present paper,

System	TA
Baseline Li et al. 2012	79.77
NoTextFeats	74.61
NoTextFeats + Best Group Comb (token)	79.56
NoTextFeats + Best Group Comb (type)	81.94*
Token-level features	
Best Gaze Group (Basic)	80.42*
Best Gaze-Only Comb (Basic+Late)	80.45*
Best Single Group (NoGazeDun)	80.61*
Best Group Comb (All)	81.00*
Type-averaged features	
Best Gaze Group (Basic)	81.28*
Best Gaze-Only Comb (Basic+Late)	81.38*
Best Group (NoGazeDun)	82.44*
Best Group Comb (All)	82.44*

Table 6.3: Tagging accuracy for the baseline, for models with no text features and for our gaze-enriched models using type and token gaze features. Significant improvements over the baseline marked by * ($p < 10^{-3}$, McNemar's test).

we use Universal part-of-speech (UPOS) tags (Petrov, Das, and McDonald, 2011), which were obtained by automatically mapping the original Penn Treebank annotation of the Dundee Treebank to Universal tags.

6.3 TYPE-CONSTRAINED SECOND-ORDER HMM POS TAGGING

We build on the type-constrained SHMM-ME proposed by Li, Graça, and Taskar (2012). This model is an extension of the first-order maximum entropy HMM introduced by Berg-Kirkpatrick et al. (2010). Li, Graça, and Taskar (2012) derive type constraints from crowd-sourced tag dictionaries obtained from Wiktionary. Using type constraints means confining the emissions for a given word to the tags specified by the Wiktionary for that word. Li, Graça, and Taskar (2012) report a considerable improvement over state-of-the-art unsupervised POS tagging models by using type constraints. In our experiments, we use the tag dictionaries they made available[2] to facilitate comparison. Li

2 https://code.google.com/archive/p/wikily-supervised-pos-tagger/

et al.'s model was evaluated across nine languages and outperformed a model trained on the Penn Treebank tagset, as well as a models that use parallel text. We follow Li et al.'s approach, including the mapping of the Penn Treebank tags to the UPOS tags (Petrov, Das, and McDonald, 2011). Figure 6.1 shows a graphical representation of a second-order HMM.

Li et al. explore two aspects of type-constrained HMMs for unsupervised POS tagging: the use of a second-order Markov model, and the use of textual features modeled by maximum entropy emissions. They find that both aspects improve tagging accuracy and report the following results for English using Universal POS tags on the Penn Treebank: first-order HMM 85.4, first-order HMM with max-ent emissions 86.1, second-order HMM 85.0, and SHMM-ME 87.1. Li et al. employ a set of basic textual features for the max-ent versions, which encode word identity, presence of a hyphen, a capital letter, or a digit, and word suffixes of two to three letters.

6.4 EXPERIMENTS

FEATURES Based on the eye-movement data in the Dundee Corpus, we compute token-level values for 22 features pertaining to gaze and complement them with another nine non-gaze features. Word length and word frequency are known to correlate and interact with gaze features. We use frequency counts from both a large corpus (British National Corpus (BNC)) and the Dundee Corpus itself. From these corpora, we also obtain forward and backward transitional probabilities, i.e., the conditional probabilities of a word given the previous or next word.

All gaze features are averaged over the ten readers and normalized linearly to a scale between 0 and 1. We divide the set of 31 features, which we list in Table 6.1, into the following seven groups in order to examine for their individual contribution:

EARLY measures of processing such as first-pass fixation duration. Fixations on previous words are included in this group due to preview benefits. Early measures capture lexical access and early syntactic processing.

LATE measures of processing such as number of regressions to a word and re-fixation probability. These measures reflect late syntactic processing and disambiguation in general.

BASIC word-level features, e.g., mean fixation duration and fixation probability. These metrics do not belong explicitly to early or late processing measures.

REGFROM includes a small selection of measures based on regressions departing from a token. It also includes counts of long

FEATURE GROUPS	ACCURACY	Δ
All groups	81.00	
−NOGAZEBNC	80.80	−0.20
−NOGAZEDUN	80.28	−0.52*
−BASIC	80.20	−0.08
−EARLY	79.78	−0.42*
−LATE	79.53	−0.25
−REGFROM	79.24	−0.29*
−CONTEXT (Baseline)	79.77	+0.53*

Table 6.4: Results of an ablation study over feature groups on the test set on token-level features. Significant differences with previous model are marked by * ($p < 0.05$, McNemar's test).

regressions[3]. The token of departure of a regression can have syntactic relevance, e.g., in garden path sentences.

CONTEXT features of the surrounding tokens. This group contains features relating to the fixations of the words in near proximity of the token. The eye can only recognize words a few characters to the left, and seven to eight characters to the right of the fixation (Rayner, 1998). Therefore it is useful to know the fixation pattern around the token.

NOGAZEBNC includes word length and word frequency obtained from the British National Corpus, as well as forward and backward transitional probabilities. These were computed using the KenLM language modeling toolkit (Heafield, 2011) with Kneser-Ney smoothing for unseen bigrams.

NOGAZEDUN includes the same features as NOGAZEBNC, but computed on the Dundee Corpus. They were extracted using CMU-Cambridge language modeling toolkit.[4]

SETUP The Dundee Corpus does not include a standard train/development/test split, so we divided it into a training set containing 46,879 tokens/1,896 sentences, a development set containing 5,868 tokens/230 sentences, and a test set of 5,832 tokens/241 sentences.

To tune the number of expectation maximisation (EM) iterations required for the SHMM-ME model, we ran several experiments on the development set using 1 through 50 iterations. The result is fairly consistent for both the baseline (the original model of Li, Graça, and

3 defined as saccades going further back than w_{i-2}

4 http://www.speech.cs.cmu.edu/SLM/toolkit.html

Taskar (2012)) and the full model (which includes all feature groups in Table 6.1). Tagging accuracy as a function of number of iterations is graphed in Figure 6.2. The best number of iterations on the full model is five, which we will use for the remaining experiments.

We perform a grid search over all combinations of the seven feature groups, using five EM iterations for training, evaluating the resulting models on token-level features of the development set. We observe that the best single feature group is NoGazeDUN, the best single group of gaze features is Basic, the best gaze-only group combination is basic-Late and the best group combination is obtained by including all seven feature groups. Using all feature groups outperforms any individual feature group on development data. The performance of all the individual groups and of the best group combinations can be seen in Table 6.2. We run experiments on the test set and report results using the best single group (NoGazeDun), the best single gaze group (Basic), the best gaze-only group combination (Basic-Late) and the best group combination (all features).

Following Barrett and Søgaard (2015a), we contrast the token-level gaze features with features aggregated at the type level. Type-level aggregation was used by Barrett and Søgaard (2015a) for supervised POS tagging: A lexicon of word types was created and the features values were averaged over all occurrences of each type in the training data.

As our baseline, we train and evaluate the original model proposed by Li, Graça, and Taskar (2012) on the train-test split described above, and compare it to the models that make use of eye-tracking measures.

To get an estimate of the effect of the textual features of Li et al., we train a model without these features, labeled NoTextFeats. We also augment this model with the best combination of feature groups.

6.5 RESULTS

The main results are presented in Table 6.3. We first of all observe that both type- and token-level gaze features lead to significant improvements over Li, Graça, and Taskar (2012), but type-level features perform better than token-level. We observe that the best individual feature group, NoGazeDUN, performs better than the best individual gaze feature group, Basic and the best gaze-only feature group, Basic+Late. This is true on both type and token-level. Using the best combination of feature groups (All features) works best for both type- and token-level features. Also when excluding the textual feature model gaze helps and type-level features also work better than token-level here.

A feature ablation study (see Table 6.4) supports the hierarchical ordering of the features based on the development set results (see Table 6.1).

6.6 RELATED WORK

The proposed approach continues the work of Barrett and Søgaard (2015a) by augmenting an unsupervised baseline POS tagging model instead of a supervised model. Our work also explores the potentials of token-level features. Zelenina (2014) is the only work we are aware of that uses gaze features for unsupervised POS tagging. Zelenina (2014) employs gaze features to re-rank the output of a standard unsupervised tagger. She reports a small improvement with gaze features when evaluating on the Universal POS tagset, but finds no improvement when using the Penn Treebank tagset.

6.7 DISCUSSION

The best individual feature group is NoGazeDUN, indicating that just using word length and word frequency, as well as transitional probabilities, leads to a significant improvement in tagging accuracy. However, performance increases further when we add gaze features, which supports our claim that gaze data is useful for weakly supervising POS induction.

Type-level features work noticeably better than token-level features, suggesting that access to eye-tracking data at test time is not necessary. On the contrary, our results support the more resource-efficient set-up of just having eye-tracking data available at training time. We assume that this finding is due to the fact that eye-movement data is typically quite noisy; averaging over all tokens of a type reduces the noise more than just averaging over the ten participants that read each token. Thus token-level aggregation leads to more reliable feature values.

Our finding that the best model includes all groups of gaze features, and that the best gaze-only group combination works better than the best individual gaze group suggest that different eye-tracking features contain complementary information. A broad selection of eye-movement features is necessary for reliably identifying POS classes.

6.8 CONCLUSIONS

We presented the first study of weakly supervised part-of-speech tagging with eye-tracking data, using a type-constrained SHMM-ME. We performed experiments adding a broad selection of eye-tracking features at training time (type-level features) and at test time (token-level features). We found significant improvements over the baseline in both cases, but type-averaging worked better than token-level features. Our results indicate that using traces of human cognitive processing, such as the eye-movements made during reading, can be used to augment NLP models. This could enable us to bootstrap

better POS taggers for domains and languages for which manually annotated corpora are not available, in particular once eye-trackers become widely available through smartphones or webcams (Skovsgaard, Hansen, and Møllenbach, 2013; Xu et al., 2015).

ACKNOWLEDGMENTS

This research was partially funded by the ERC Starting Grant LOWLANDS No. 313695, as well as by Trygfonden.

7

CROSS-LINGUAL TRANSFER OF CORRELATIONS BETWEEN POS AND GAZE FEATURES

ABSTRACT

Several recent studies have shown that eye movements during reading provide information about grammatical and syntactic processing, which can assist the induction of NLP models. All these studies have been limited to English, however. This study shows that gaze and POS correlations largely transfer across English and French. This means that we can replicate previous studies on gaze-based POS tagging for French, but also that we can use English gaze data to assist the induction of French NLP models.

7.1 INTRODUCTION

The eye movements during normal, skilled reading are known to reflect the processing load associated with reading. Recently, eye movement data has been integrated into natural language processing models for weakly supervised POS induction (Barrett et al., 2016), sentence compression (Klerke, Goldberg, and Søgaard, 2016), supervised POS tagging (Barrett and Søgaard, 2015a), and supervised parsing (Barrett and Søgaard, 2015b).

Barrett et al. (2016) used eye movements from the English portion of a large eye tracking corpus, the Dundee corpus (Kennedy, Hill, and Pynte, 2003), for weakly supervised POS induction for English, obtaining significant improvements over a baseline without gaze features. They used a second-order hidden Markov Model, which was type-constrained by Wiktionary dictionaries for their experiments. These results suggest an approach to weakly supervised POS induction using only a dictionary and eye movement data. Such an approach would be applicable for low-resource languages, for which it is difficult to find professional annotators.

The present study further explores to which extent native readers' processing of POS generalizes across related languages. We use a similar model as Barrett et al. (2016), but perform cross-lingual experiments with both the French and the English portion of the Dundee Corpus.

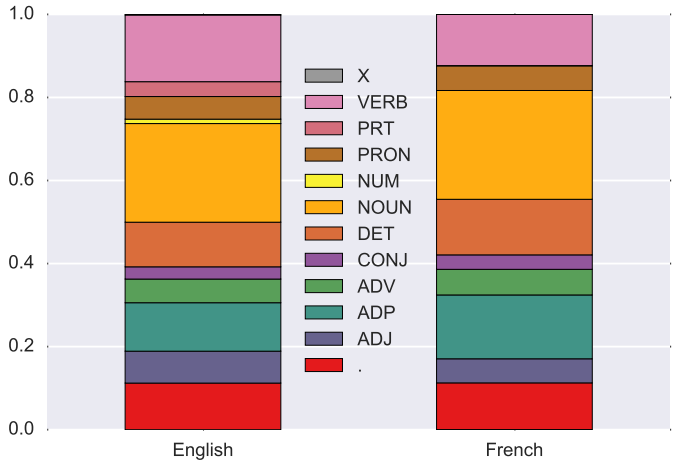


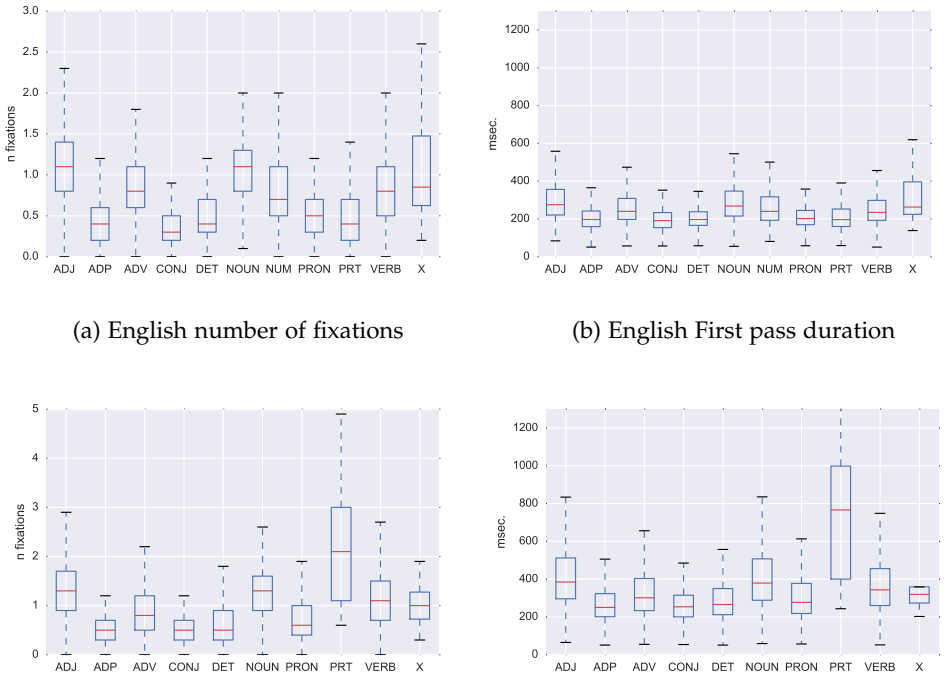
Figure 7.1: Distribution of POS in the English and French training sets.

CONTRIBUTION This is to the best of our knowledge the first study to explore how the eye movements of native readers that inform POS models generalize from one language to another. We also introduce a new resource for studying the relation between grammatical class and eye movements in French: we provide POS annotation for most of the French Dundee Corpus by aligning it with the morphosyntactic annotation of the French Treebank (Abeillé, Clément, and Toussenel, 2003).

7.2 DATA PREPARATION

The data used for this experiment is the English and French portions of the Dundee Corpus (Kennedy, Hill, and Pynte, 2003). The Dundee Corpus is the largest available eye movement corpus by token count. For English and French, 10 native speakers of each language read 20 newspaper articles from either *The Independent* (English) or *Le Monde* (French). The corpus comprises around 50,000 tokens per language.

For both the English and the French part of the Dundee Corpus, the original tokenization follows the visual units of the text, and contractions and punctuation are attached to the word whose visual unit they belong to. For instance, *s'entendre* or *rappelle-t-il* are one token in the French Dundee Corpus but two and five, respectively, in the French Treebank. In the English Dundee Corpus, *don't!* is one token, but three in the Dundee Treebank. As a result, eye movement measures are only available for the entire visual unit. We address this issue by duplicating the eye movement measures for all treebank tokens that comprise a Dundee token (i.e., a visual unit). This is the same approach Barrett et al. (2016) used. As a result, the number of tokens increases in the POS tagged version of the Dundee Corpus;



(a) English number of fixations

(b) English First pass duration

(c) French number of fixations

(d) French first pass duration

Figure 7.2: Two reading measures across POS class computed on the English and French training sets.

also, some tokens are associated with eye movement measures that reflect the processing of several tokens.

For English, the treebank tokenization leads to 13.8% increase of tokens to 58,599 tokens. For French, the treebank tokenization leads to an 17.7% increase on token count to 56,683 tokens. For the English training set, 76% of all Dundee Corpus tokens are mapped to one treebank token. The same goes for 62% of the Dundee Corpus tokens for French.

7.2.1 *English*

The Dundee Treebank (Barrett, Agić, and Søgaard, 2015) is a recent manual, syntactic annotation layer for the English portion of the Dundee Corpus following the Universal Dependency formalism. For evaluation, we use the POS labels from this resource. We mapped the Penn Treebank tagset used in the Dundee Treebank automatically to the Universal POS tag set (Petrov, Das, and McDonald, 2011).

The split into training, development, and test set for the English Dundee corpus is identical to the splits used by Barrett et al. (2016), with 80% of the tokens for training and 10% of the tokens for development and testing, respectively, without splitting up sentences. This

split results in 46,879 tokens in 1,896 sentences for training, 5,868 tokens in 230 sentences for development, and a test set of 5,832 tokens in 241 sentences.

7.2.2 *French*

The text for the French part of the Dundee Corpus is originally from the French Treebank version 1.4 (Abeillé, Clément, and Toussenel, 2003) and we re-aligned the two corpora for this experiment. We first manually identified the relevant subset of the French Treebank (which is discontinuous). A small part (2,518 tokens equivalent of 5.31% of the French Dundee tokens) of the Dundee Corpus could not be found by manual search in the French Treebank and was therefore omitted from the experiment. Only entire sentences were removed. The morphosyntactic annotation of the French Treebank was semi-manually aligned with the Dundee Corpus by a set of heuristic rules and by manually fixing all exceptions. Due to tokenization inconsistencies in both the French Treebank and the Dundee Corpus, manual intervention was required.

For French there are some treebank tokens with no token string, only POS lemma etc. For example, *du* should be split into *de* and *le*, but in some instances the token string for *le* is missing. These missing tokens were omitted from this experiment.

The French Dundee Corpus does not come with a train-development-test split. We use a similar approach as for English, with the first 80% of the tokens for training, the next 10% of the tokens for development and the last 10% for testing. No sentences were split into separate sets. That results in 43,383 tokens in 1,585 sentences for training, 5,407 tokens in 240 sentences for development, and 5,444 tokens in 178 sentences for testing.

The tagset of the French Treebank was automatically mapped to the Universal POS tag set (Petrov, Das, and McDonald, 2011). We make the aligned, morphosyntactic annotation for the French Dundee Corpus available at `https://bitbucket.org/lowlands/release`.

7.2.3 *Reading differences between English and French*

This section discusses the results of existing studies comparing reading in French and English. The two main studies used the two Dundee corpora for their analysis.

Pynte and Kennedy (2006) compared the eye movements of the French and English Dundee corpus to explore local effects (e.g., word frequency, word length, local context) and global effects (e.g., predictability, reading strategy, inspection strategy) on five eye movement metrics.

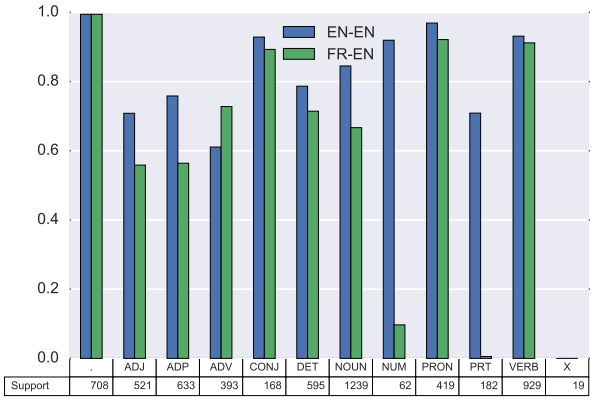
	− SUFFIX FEATS			+ SUFFIX FEATS		
TR-TE	NO GAZE	TOKEN	TYPE	NO GAZE	TOKEN	TYPE
Development set accuracy						
EN-EN	77.44	80.01	**83.38**	80.21	81.46	**83.86**
FR-EN	**73.16**	72.92	72.92			
FR-FR	82.45	83.08	**86.55**	83.39	84.11	**87.52**
EN-FR	79.38	80.86	**80.97**			
Test set accuracy						
EN-EN	76.49	78.49*	**82.14***	80.37	80.60	**83.25***
FR-EN	71.38	71.39	**71.58**			
FR-FR	81.30	82.27*	**85.03***	83.16	83.30*	**86.22***
EN-FR	78.34	79.83*	**79.92***			

Table 7.1: Accuracy on development and test set for type-and token-level experiments. Best condition per experimental set-up per language combination in bold. *) For test set results: $p < 0.001$ according to mid-p McNemar test when compared to baseline.

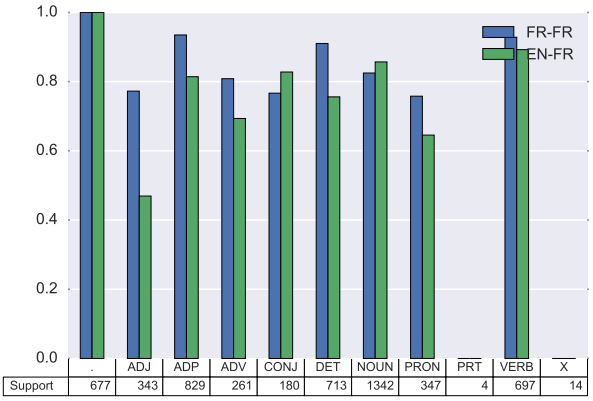
They first of all noted that French was read slower than English with more and longer fixations. This effect is significant and is even more pronounced for long words and there are also significantly more re-fixations for French compared to English. Kennedy and Pynte (2005) argue that re-fixations reflect the most crucial difference between French and English. Besides being an obvious difference in the processing of the target word, more re-fixations also enhance preview of the next word. Pynte and Kennedy (2006) report that participants of the English and French experiments were matched (though not on which factors) and that the procedure, including calibration technique, equipment, control software, instructions, and data-reduction software, were identical across language, though the French data was collected in Aix-en-Provence, France and the English data in Dundee, UK. Therefore they ascribed this difference to the text itself. Even though they found that French words (5.2 characters) are on average longer than English ones (4.7 characters), there are more two-letter words in French (19.7%) than in English (17.2%). Therefore Kennedy and Pynte (2005) suggest that the reading difference is due the distribution of information across the letters of a given words, which is different across these two languages. For example, in French, terminal accents, case markers, and gender and tense marking convey crucial morphological information. This is in line with their finding that eye movements in the English part of the Dundee Corpus were

more sensitive to the length of the next word, whereas French showed equivalent effects of the informativeness of the word-initial trigram.

Overall, Kennedy and Pynte (2005) and Pynte and Kennedy (2006) conclude that the English and French inspection strategies are remarkably similar, which is the same conclusion Sparrow, Miellet, and Coello (2003) made when testing the English EZ reader model on another eye movement corpus of 134 words of French. Kennedy and Pynte (2005) provide an analysis of the statistical differences between English and French, but besides re-fixations being more frequent in French, they seem to conclude that the reading is in many respects similar, which is also supported by their choice of mainly analyzing French and English jointly.



(a) EN-EN and FR-EN



(b) FR-FR and EN-FR

Figure 7.3: Accuracy on development set for all POS classes.

The treebank annotation includes sentence boundaries, which makes it possible to compare the length and the complexity of the sentences for both languages. We find that the average sentence length of the English training set is 24.7 tokens (SD 13.1). For French it is 28.7 tokens (SD 17.8). Sentence length was not considered by Kennedy

and Pynte (2005) and Pynte and Kennedy (2006). A consequence of longer sentences is that reading difficulty increases. The Coleman-Liau index (Coleman and Liau, 1975) is 10.38 for the English training set and 12.98 for the French.[1] This could stem from different writing styles in *Le Monde* and *The Independent* or a biased sampling of articles.

The conclusion can go no further than to say that French and English readers *can* display a more or less similar inspection strategy when reading text under matched conditions. Some effects, e.g., the fact that word-initial trigrams are more important for fixation durations in French than in English, could be due to cross-lingual differences in the spelling of the two languages, leading to re-fixations in order to increase preview. But slower reading could also be partly due to the presence of more difficult texts in the French corpus. See Section 7.7 for a further discussion on grammatical processing differences across languages.

7.2.3.1 *Comparing reading of POS for English and French*

The statistics presented in the following section were computed on the French and English training sets and extends the comparison of Section 7.2.3 with respect to POS We show that the POS classes are overall read similarly across the two languages with few exceptions due to systematic biases.

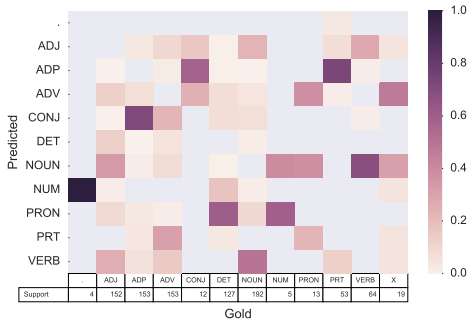
Figure 7.1 shows the distribution of POS classes in the English and French data. The biggest differences are that there are no NUM tags in French. This is due to the annotation scheme and our automatic mapping, in which no tags map to NUM. There are also very few particles in the French data compared to English.

Figure 7.2 shows boxplots for two different reading metrics: number of fixations and first pass duration, across POS class for English and French. The first pass duration is the sum of fixation durations for a token in the first pass through the text. This measure is said to encompass early syntactic and lexical processing. The number of fixations encompasses re-fixations and regressions to a token and reflects later syntactic and semantic processing.

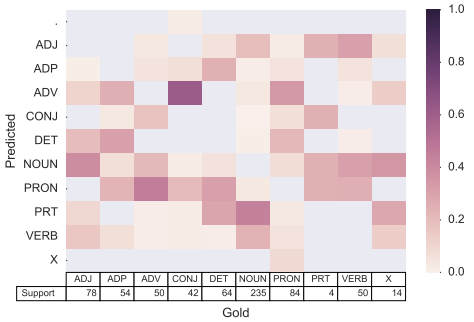
Note that punctuation is almost always glued to a word and any eye movements on a punctuation will mainly–if not solely–reflect the processing of the other token. Therefore punctuation is excluded from Figure 7.2.

When comparing Figure 7.2d and Figure 7.2b, we can confirm the findings of Pynte and Kennedy (2006) that fixations are generally longer in the French portion than in the English portion of Dundee. Average gaze duration in the training set is 236 ms for English and 303 ms for French.

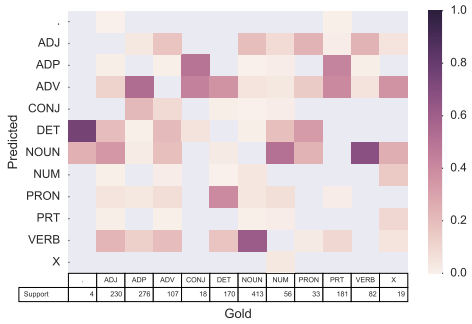
1 calculated using http://www.online-utility.org/



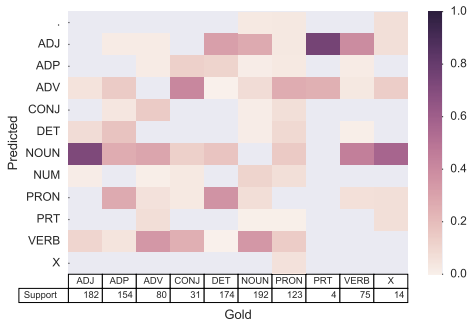
(a) EN-EN



(b) FR-FR



(c) FR-EN



(d) EN-FR

Figure 7.4: Erroneous predictions per gold POS for all combinations of training and testing language on development set.

It can be seen from Figure 7.2 that the measures differ across POS for most classes in an intuitive way. For instance, POS classes of short, frequent, closed-class words such as CONJ, ADP, PRON and DET get fewer and shorter fixations than, e.g., NOUN, VERB, ADJ, and ADV. This seems to be consistent across the two languages, and is in line with a similar analysis for English (Barrett and Søgaard, 2015a) for a smaller data set of naturally occurring text from five different domains.

The PRT category seems to be an exception. In French, PRT seems to require extensive early and late processing. Remember from Figure 7.1 that there are more PRTs for English (3.6%) and fewer for French (0.05%). The sets of PRT words for the two languages reveal a systematic bias in the annotation scheme or automatic mapping. For the French training set, the set of PRTs is {*vice-*, *pseudo-*, *post-*, *contre-*, *anti-*, *non-*, *quasi-*, *soviéto-*, *supra-*, *néo-*, *inter-*}. For English it is {*off*, *down*, *To*, *about*, *on*, *in*, *over*, *around*, *back*, *up*, *out*, *to*, *away*, *'*, *'s*}. French particles are therefore always at least two-token visual units that seem to be quite infrequent as well as long, whereas English particles are short and frequent.

7.3 FEATURES

For our weakly supervised POS tagging experiments, we use 22 gaze features that measure both early processing and late processing. They are equivalent to the 22 gaze features used by Barrett et al. (2016). Early processing measures are said to reflect different aspects of early syntactic and semantic processing and include first pass duration and first fixation duration. Late processing measures reflect, e.g., late syntactic and semantic integration (Rayner, 1998). Examples are number and duration of regressions going to a word, as well as the total reading time for a word.

Non-gaze features are usually included in eye movement models, because they explain a lot of the variance in fixation durations. Word frequency and word length together have been found to explain 69% of the variance in the mean gaze duration (Carpenter and Just, 1983). Like Barrett et al. (2016), we use word length, log word frequencies from a big corpus and log word frequencies from the Dundee training set for the target word, and the previous and next words. From the Dundee training set, we also extract the forward and backward transitional probability, i.e., the conditional probabilities for a word given the next or previous word. Our non-gaze features are almost equivalent to Barrett et al. (2016). The only difference is that they also used forward and backward transitional probabilities from a big corpus.

The big corpus log frequencies were obtained from the British National Corpus (BNC)[2] for English, extracted with KenLM (Heafield,

2 http://www.natcorp.ox.ac.uk

2011) and Lexique[3] for French. The Dundee log frequencies were calculated on the respective training sets using CMU Language Modeling Toolkit[4] with Witten-Bell smoothing.

In total we have 29 features. All features are first averaged over all 10 readers of the corpus, then scaled to a value between 0 and 1 by min-max scaling. The best model of the feature ablation study of Barrett et al. (2016) used all features, which suggests that grammatical processing of a broad set of POS categories is reflected across many features and need non-gaze features as well.

7.4 EXPERIMENT

We replicate the experimental setup of Barrett et al. (2016), which used the best model from Li, Graça, and Taskar (2012), a second-order hidden Markov model with maximum entropy emissions (SHMM-ME), constrained by Wiktionary tags such that emissions are confined to the allowed POS tags of the Wiktionary given that the token exists in the Wiktionary. Li, Graça, and Taskar (2012) report considerable improvements from the Wiktionary contraint when comparing to unsupervised methods.

The second-order model includes transition probabilities from the antecedent state like a first order model (Berg-Kirkpatrick et al., 2010) as well as from the second-order antecedent state.

We use the original implementation of Li et al. and we also include a subset of their word-level features, viz., four features detecting hyphens, numerals, punctuation and capitalization. We leave out the three suffix features from Li et al.'s basic feature model, as these features do not transfer across languages. These features were also included by Barrett et al. (2016).

We use the English Wiktionary dumps made available by Li et al.[5] The French Wiktionary dump is from Wisniewski et al. (2014) and does not include any punctuation. We therefore augment it with all punctuation entries from the English Wiktionary. Furthermore, tokens for the tag ADP are completely missing from the French Wiktionary, and the tokens for the class DET were sparse. We therefore add all examples of DET and ADP from the French training set to the French Wiktionary.

For the cross-lingual experiments, we use the union of the French and the English Wiktionary dictionaries.

Barrett et al. (2016) used Li et al.'s model for weakly supervising POS induction with gaze features for English, and performed model tuning and feature ablation. We use their best hyper-parameter setting, i.e., five EM iterations, as well as the best feature combination:

3 http://www.lexique.org
4 http://www.speech.cs.cmu.edu/SLM/toolkit.html
5 https://code.google.com/archive/p/wikily-supervised-pos-tagger/

Metric	Cosine sim
n refixations	0.6318
First pass duration	0.8480
Re-read probability	0.8489
n fixations	0.9097
Total fixation duration	0.9217
n regressions to	0.9354
n long regressions from	0.9375
Total duration of regressions from	0.9377
Total duration of regression to	0.9385
n regressions from	0.9404
n long regressions to	0.9644
Fixation probability	0.9795
w-1 fixation duration	0.9839
w+1 fixation duration	0.9934
w-1 fixation probability	0.9947
w+2 fixation duration	0.9961
w+1 fixation probability	0.9967
w-2 fixation probability	0.9975
w+2 fixation probability	0.9986
First fixation duration	0.9992
Mean fixation duration	0.9992
w-2 fixation duration	0.9992

Table 7.2: Cosine similarity between POS-averaged French and English train set gaze vectors across gaze features. Sorted by similarity.

all features. Following Barrett et al. (2016), we try token-level and type-level features. For the token-level experiments, each token is represented by its feature vector. For the type-level experiments, each token is represented by an average of the feature vectors for all occurrences of the lower-cased word type of the training set.

7.5 RESULTS

The tagging accuracy for all combinations of training and testing language on the development set and the test set can be seen in Table 7.1.

For all conditions, type-level features work better than token-level, though the type-level improvement over the baseline is not significant for FR-EN.

The English monolingual condition plus suffix is almost equivalent to the best model in Barrett et al. (2016). The only difference is the two missing non-gaze features described in Section 7.3. On the test set, they report a baseline accuracy of 79.77, a token-level accuracy of 81.00, and a type-level accuracy of 82.44, which is in line with our results. We observe that the suffix features seem to help in the monolingual conditions. For monolingual conditions, we confirm that type-level gaze-features and token-level ones outperform the baseline. These differences are significant, except for the EN-EN token-level plus suffix condition.

FR-FR POS tagging seems to be a slightly an easier task than EN-EN POS tagging, achieving overall higher accuracies.

The cross-lingual conditions generally achieve lower performance than the monolingual. When training on English and testing on French, both token-level and type-level conditions are significantly better than baseline.

7.6 ERROR ANALYSIS

There are – as expected – more errors when using cross-lingual gaze data. This section will explore these errors by comparing the predictions of the cross-lingual experiments with the predictions of the monolingual experiments. All analysis is on the development set output of the type-level models. We compare them to the output of the type-level monolingual models.

Figure 7.3 shows accuracy scores per POS class comparing experiments with same test set. The accuracy of punctuations is due to the basic feature model and the Wiktionary constraints–not the eye movement measures. PRT and NUM are real challenges for FR-EN compared to EN-EN. This can be assumed to be due to the different use of the PRT tag and the missing NUM class in the French dataset described in Section 7.2.3.1. ADJ also seems like a cross-lingual challenge, though harder when trained on English and tested on French than the other way around.

Figure 7.4 shows the erroneous predictions per gold POS tag, allowing us to compare error types across experiments. When comparing Figure 7.4a and Figure 7.4c, both evaluated on English, most classes seem to have almost the same set of misclassified labels though for some labels in different magnitude or ratio depending on whether they are trained on English or French. The main differences are: when trained on French, ADP and ADJ are generally more often misclassified, ADP is not mainly misclassified as CONJ, but more often as ADV, DET is also misclassified as VERB and ADV, PRT is misclassified as ADV and not mainly as ADP.

When comparing Figure 7.4b and Figure 7.4d, both evaluating on French, we also find that for many of the POS classes, the misclassifi-

cations are of the same type, though different in magnitude or ratio. The main differences we observe when training on English are: ADJ is mainly misclassified as NOUN instead of ADP, ADV, DET, NOUN, and PRT; ADV is misclassified as VERB; DET is never misclassified as PRT, but more often as NOUN and ADJ; and NOUN is rarely misclassified as PRT. The last error probably has to do with the long gaze durations for PRT in the French data (resembling gaze durations of NOUNs) opposed to the short gaze durations of English PRT.

Table 7.2 shows the cosine similarity between the English and French POS-averaged gaze vectors from the train set for all gaze features. This gives information about which gaze feature averages differ between French and English POS Pynte and Kennedy (2006) found that French had more re-fixations than English, which is reflected in the table. Measures correlating with re-fixations like re-read probability, number of fixations, and total fixation duration are naturally also different across languages. First pass duration is not directly correlated with number of re-fixations, and must be considered an distinct pattern.

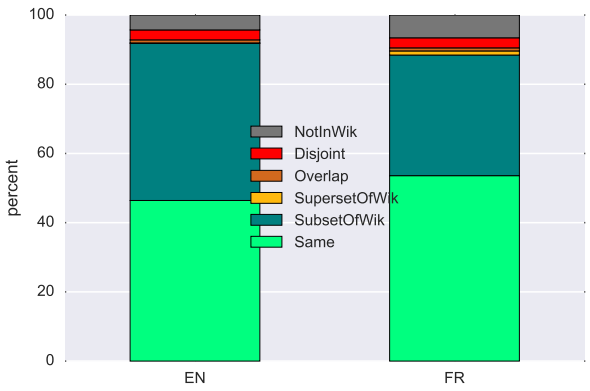


Figure 7.5: Development set word type lookup in Wiktionary for English and French: the percentage of word types assigned a set of tags that is either: identical to, a subset of, a superset of, overlapping with, disjoint with, or not in the Wiktionary.

7.6.1 *Wiktionary agreement*

Figure 7.5 shows the word types for the English and French development set according to their representation in the respective monolingual Wiktionary. This figure is inspired by Li, Graça, and Taskar (2012). For English, more POS types agree with the Wiktionary (Same or SubsetOfWik) than for French. We also computed token-level accuracies, where a tag licensed by Wiktionary counts as correct. For the French development set, this maximum dictionary accuracy is 0.95, whereas for English it is 0.92.

7.7 DISCUSSION

We presented four experiments with POS induction using gaze data in a monolingual and cross-lingual setup with a second-order hidden Markov model. Our experiments confirm the main conclusion from Barrett et al. (2016), viz., that type-level gaze vectors improve POS induction. We replicated their result for English and report the same finding for French as well as for French when trained on English gaze vectors.

It is difficult to determine how much the relatedness of the French and English languages is responsible for the ability of the model to generalize cross-lingually. The psycholinguistic literature does not reveal how different POS categories are processed across languages; most experimental work in the literature studies single phenomena in one language. For instance, in reaction time studies of lexical decision tasks it has been found that the processing of English plural and singular nouns is influenced by surface frequency only[6] (Sereno and Jongman, 1997), whereas for Dutch (Baayen, Dijkstra, and Schreuder, 1997) and French (New et al., 2004), the lexical processing of singular and plural nouns is influenced by the base frequency[7]. The English data thus support a full-storage cognitive model, whereas the French and the Dutch data support the Parallel Dual-Route model where a word is processed as segments in parallel with whole word processing. These results suggest that nouns are processed differently in the brain for native speakers of different languages. This means that our results may not generalize to other combinations of languages and in the specific case of nouns it suggests that Dutch and French nouns are processed more similarly than French and English.

7.8 CONCLUSION

This is, to the best of our knowledge, the first study to explore whether gaze features generalize from one language to another for a broad set of syntactic categories. We used a type-constrained second-order HMM for monolingual and cross-lingual POS induction on the English and French portions of the Dundee eye tracking corpus. We experimented with both token-level and type-level features and confirmed that type-level gaze features improve monolingual POS induction for both English and French. We also showed that type-level gaze features significantly improve POS induction for French, even when the model is trained on English gaze vectors.

6 the token frequency of a word form
7 the sum of the frequencies of all inflections of a word

8

UNSUPERVISED INDUCTION OF LINGUISTIC CATEGORIES WITH RECORDS OF READING, SPEAKING, AND WRITING

ABSTRACT

When learning POS taggers and syntactic chunkers for low-resource languages, different resources may be available, and often all we have is a small tag dictionary, motivating type-constrained unsupervised induction. Even small dictionaries can improve the performance of unsupervised induction algorithms. This paper shows that performance can be further improved by including data that is readily available or can be easily obtained for most languages, i.e., eye-tracking, speech, or keystroke logs (or any combination thereof). We project information from all these data sources into shared spaces, in which the union of words is represented. For English unsupervised POS induction, the additional information, which is not required at test time, leads to an average error reduction on Ontonotes domains of 1.5% over systems augmented with state-of-the-art word embeddings. On Penn Treebank the best model achieves 5.4% error reduction over a word embeddings baseline. We also achieve significant improvements for syntactic chunk induction. Our analysis shows that improvements are even bigger when the available tag dictionaries are smaller.

8.1 INTRODUCTION

It is a core assumption in linguistics that humans have knowledge of grammar and that they use this knowledge to generate and process language. Reading, writing, and talking leave traces of this knowledge and in psycholinguistics this data is used to analyze our grammatical competencies. Psycholinguists are typically interested in falsifying a specific hypothesis about our grammatical competencies and therefore collect data with this hypothesis in mind. In NLP, we typically require big, representative corpora. NLP usually has induced the models from expensive corpus annotations by professional linguists, but recently, a few researchers have shown that data traces from human processing can be used directly to improve NLP models (Barrett et al., 2016; Klerke, Goldberg, and Søgaard, 2016; Plank, 2016a).

In this paper, we investigate whether unsupervised POS induction and unsupervised syntactic chunking can be improved using human text processing traces. We also explore what traces are beneficial, and

how they are best combined. Our work supplements psycholinguistic research by evaluating human data on larger scale than usual, but more robust unsupervised POS induction also contributes to NLP for low-resource languages for which professional annotators are hard to find, and where instead, data from native speakers can be used to augment unsupervised learning.

We explore three different modalities of data reflecting human processing plus standard, pre-trained distributional word embeddings for comparison, but also because some modalities might fare better when combined with distributional vectors. Data reflecting human processing come from reading (two different eye-tracking corpora), speaking (prosody), and typing (keystroke logging). We test three different methods of combining the different word representations: a) canonical correlation analysis (CCA) (Faruqui and Dyer, 2014b) and b) singular value decomposition and inverted softmax feature projection (SVD+IS) (Smith et al., 2017) and c) simple concatenation of feature vectors.

CONTRIBUTIONS We present experiments in unsupervised POS and syntactic chunk induction using multi-modal word representations, obtained from records of reading, speaking, and writing. Individually, all modalities are known to contain syntactic processing signals, but to the best of our knowledge, we are the first to combine them in one model. Our work extends on previous work in several respects:

A. We compare using data traces from gaze, speech, and keystrokes.

B. We consider three ways of combining such information that do not require access to data from all modalities for all words.

C. While some previous work assumed access to gaze data at test time, our models do not assume access to any modalities at test time.

D. We evaluate how much the additional information helps, depending on the size of the available tag dictionary.

E. While related work on keystrokes and prosody focused on a single feature, all our word representations are multi-dimensional and continuous.

8.2 RELATED WORK

EYE-TRACKING data reflect the eye movements during reading and provide millisecond-accurate records of the readers fixations. It is well established that the duration of the fixations reflect the processing load of the reader (Rayner, 1998). Words from closed word classes are usually fixated less often and for shorter time than words from open

word classes (Rayner and Duffy, 1988). Psycholinguistics, however, is generally not interested in covering all linguistic categories, and psycholinguists typically do not study corpora, but focus instead on small suites of controlled examples in order to explore human cognition. This is in contrast with NLP. Some studies have, however, tried to bridge between psycholinguistics and NLP. Demberg and Keller (2008) found that eye movements reflected syntactic complexity. Barrett and Søgaard (2015a) and Barrett and Søgaard (2015b) have tried to – respectively – predict a full set of syntactic classes and syntactic functions across domains in supervised setups. Barrett et al. (2016), which is the work most similar to ours, used eye-tracking features from the Dundee Corpus (Kennedy, Hill, and Pynte, 2003), which has been augmented with POS tags by Barrett, Agić, and Søgaard (2015). They tried for POS induction both on token-level and type-level features. They found that eye-tracking features significantly improved tagging accuracy and that type-level eye-tracking features helped more than token-level. We use the same architecture as Barrett et al. (2016).

KEYSTROKE LOGS also reflect the processing durations, but of writing. Pauses, burst and revisions in keystroke logs are used to investigate the cognitive process of writing (Baaijen, Galbraith, and Glopper, 2012; Matsuhashi, 1981). Immonen and Mäkisalo (2010) found that for English-Finnish translation and monolingual Finnish text production, predicate phrases are often preceded by short pauses, whereas adpositional phrases are more likely to be preceded by long pauses. Pauses preceding noun phrases grow with the length of the phrase. They suggest that the difference is explained by the fact that the processing of the predicate begins before the production of the clause starts, whereas noun phrases and adpositional phrases are processed during writing. Pre-word pauses from keystroke logs have been explored with respect to multi-word expressions (Goodkind and Rosenberg, 2015) and have also been used to aid shallow parsing (Plank, 2016a) in a multi-task bi-LSTM setup.

PROSODIC FEATURES provide knowledge about how words are pronounced (tone, duration, voice etc.). Acoustic cues have already been used to improve unsupervised chunking (Pate and Goldwater, 2011) and parsing (Pate and Goldwater, 2013). Pate and Goldwater (2011) cluster the acoustic signal and use cluster label as a discrete feature whereas Pate and Goldwater (2013) use a quantized word duration feature.

Plank (2016a) and Goodkind and Rosenberg (2015) also used a single keystroke feature (keystroke pre-word pause) and the former study also discretized the feature. Our work, in contrast, uses acoustic and keystroke features as multi-dimensional, continuous word representations.

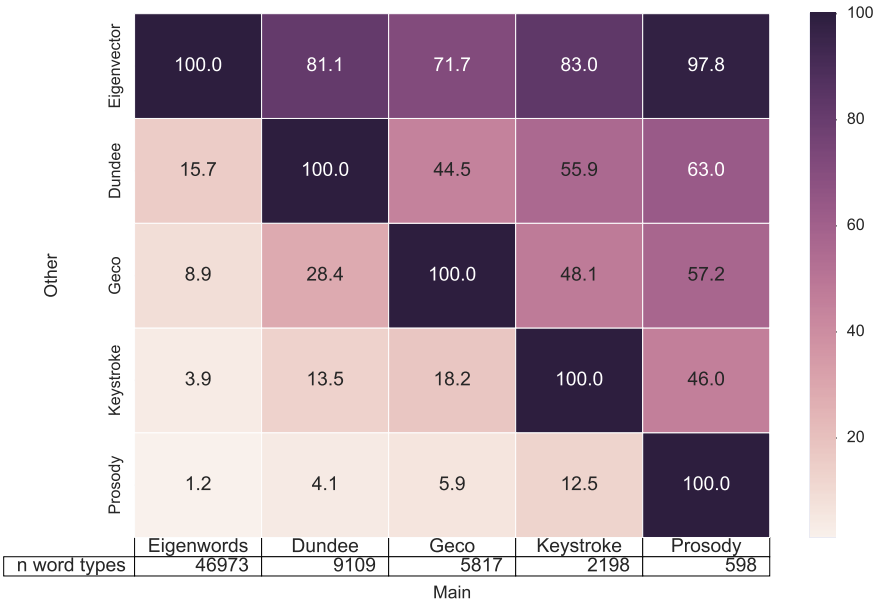


Figure 8.1: The percentage of overlapping word types for pairs of modalities. Overlapping words are used for projecting word representations into a shared space. Read column-wise. E.g. when combining eigenwords and prosody, only 1.2% of the 46973 eigenvector word types are overlapping (bottom left), and 97.8% of the 598 prosody word types are overlapping (top right).

8.3 MODALITIES

In our experiments, we begin with five sets of word representations: prosody, keystroke, gaze as recorded in the GECO corpus, gaze as recorded in the Dundee corpus, as well as standard, text-based word embeddings from eigenwords. See below for details and references. All modalities except the pre-trained word embeddings reflect human processing of language. For all modalities, we use type-level-averaged features of lower-cased word types.

The choice of using type-averaged features is motivated by Barrett et al. (2016), who tried both token-level and type-averaged eye-tracking features for POS induction and found that type-level gaze features worked better than token-level. Type-averaged features also have the advantage of not relying on access to the auxillary data at test time. Type-level averages are simply looked up in an embedding file for all previously seen words. On the other hand, type-level features obviously do not represent ambiguities, e.g., *beat* as a verb and a noun separately. All our features, except log-transformed word frequencies were normalized.

We run unsupervised induction experiments for all $(2^5 - 1 = 31)$ combinations of our five data sources on the development sets to determine which data types contribute to the task. We consider three different ways of combining modalities, two of which learn a projec-

MODALITY	*n* FOUND PAIRS	WEIGH. AV. COR.
Prosody	31	0.369
Keystroke	1082	0.060
GECO	2449	-0.030
Dundee	4066	-0.035
Eigenwords	9828	0.197

Table 8.1: Results on word association norms from wordvectors.org. Correlation weighted by number of found pairs per word embedding type.

tion into a shared space using word overlap as supervision, and one simply concatenates the embedding spaces. The combination methods are further described in Section 8.4.

We list the number of word types per modality and percentage of pair-wise overlapping words in Figure 8.1. We only use existing data from native speaking participants, for reproducibility and in order not to get learner effects ie. biases introduced by non-native speakers. Section 8.3.2, Section 8.3.3, and Section 8.3.4 describe each modality in detail, and how we compute the word representations. Section 8.3.1 describes a set of basic features used in all of our experiments.

8.3.1 *Basic features*

Like Li, Graça, and Taskar (2012), we append a small set of basic features to all our feature sets: features relating to orthography such as capitalization, digits and suffix. Furthermore we append log word frequency and word length. Word frequencies per million are obtained from BNC frequency lists (Kilgarriff, 1995). Word length and word frequency explain around 70% of the variance in the eye movement (Carpenter and Just, 1983) and are therefore also important for estimating the impact of gaze features beyond such information. Plank (2016a) used keystroke features for shallow chunking and did not find any benefit of normalizing word length by pre-word pause before typing each word, but Goodkind and Rosenberg (2015) did find a strong logarithmic relationship between word length and pre-word pause as well as between word frequency and pre-word pause.

8.3.2 *Dundee and GECO eye-tracking corpora*

We use two different eye-tracking corpora. The GECO corpus (Cop et al., 2017) and the Dundee Corpus (Kennedy, Hill, and Pynte, 2003) are the two largest eye movement corpora with respect to word count.

We use the native English part of the GECO corpus and the English part of the Dundee Corpus. The GECO corpus is publicly available[1] and the Dundee Corpus is available for research purposes.

PARTICIPANTS AND DATA The Dundee Corpus is described in Kennedy and Pynte (2005). The Dundee Corpus consists of the eye movements of 10 readers as they read the same 20 newspaper articles. For GECO, all 14 participants in the native English part read a full Agatha Christie novel. Both corpora contain $>$ 50.000 words per reader. All participants for both corpora are adult, native speakers of English and skilled readers.

SELF-PACED READING Both eye-tracking corpora reflect natural reading by making the reading self-paced and using naturally occurring, contextualized text.

FEATURES Eye movements – like most features reflecting human processing – are very susceptible to experiment-specific effects e.g. instructions and order effects such as fatigue. Furthermore, the GECO corpus has a slightly different eye movement feature set than what we have for the Dundee corpus. Therefore we treat the two eye movement corpora as two individual modalities in order to assess their individual contributions. GECO has 34 features reflecting word-based processing. Dundee has 30 word-based features that were extracted from the raw data and previously used for POS induction by Barrett et al. (2016). For GECO, we use the features that are already extracted by the authors of the corpus. Both corpora include five word-based features e.g., first fixation duration (which is a measure said to reflect early syntactic and semantic integration), total fixation time and fixation probability. The Dundee Corpus has more features concerning the context words whereas GECO has pupil size and many features distinguishing the different passes over a word.

8.3.3 *Prosody*

The prosody features are described in detail in Frermann and Frank (2017) and are freely available.[2] They are derived from the Brent (Brent and Siskind, 2001) and Providence (Demuth, Culbertson, and Alter, 2006) portions of the CHILDES corpus (MacWhinney, 2000), comprising longitudinal datasets of raw speech directed to 22 children, and its transcription. Word-level speech-text alignments were obtained automatically using forced alignment. For each token-level audio snippet, a set of 88 prosody features was extracted based on a previously established feature set (Eyben et al., 2016), including

1 http://expsy.ugent.be/downloads/geco/
2 https://github.com/ColiLea/prosodyAOA

standard features derived from F0–F3 formants, spectral shape and rhythm features, intensity and MFCC features among others. Type-level prosody features were obtained as averaged token-level features for each word type.

8.3.4 *Keystroke features*

We extracted keystroke features from the publicly available data from Killourhy and Maxion (2012). This data contains key hold times and pauses of all key presses of 20 subjects as they completed transcription and free composition tasks. We only used data from the free composition part. A pause is defined by the authors as the duration from keydown to keydown. The free composition data consists of a total of 14890 typed words and 2198 word types.

For each word, we extracted the following features:

- average key hold duration of all characters associated with producing the word.
- pre-word pause
- hold duration of space key before word
- pause length of space key press pause before word
- ratio of keypresses used in the word production to length of the final word.

For each word, we also included these five features for up to 3 words before. In total, we have $5 * 4 = 20$ keystroke features. We use lower-cased word type averages, as with the other modalities.

8.3.5 *Eigenwords*

Eigenwords are standard, pre-trained word embeddings, induced using spectral-learning techniques (Dhillon, Foster, and Ungar, 2015). We used the 30-dimensional, pre-trained eigenvectors.[3]

8.3.6 *Preliminary evaluation*

Our application of these word representations and their combinations is unsupervised POS and syntactic chunk induction, but before presenting our projection methods in Section 8.4 and our experiments in Section 8.5, we present a preliminary evaluation of the different modalities using word association norms.

3 http://www.cis.upenn.edu/~ungar/eigenwords/

Table 8.1 shows the weighted correlation between cosine distances in the representations and the human ratings in the word association norm datasets available at wordvectors.org (Faruqui and Dyer, 2014a). Eigenwords, not surprisingly, correlates better than the representation based on processing data – with the exception of prosody. The correlation with prosody is non-significant, however, because of the small sample size.

8.4 COMBINING DATASETS

We now have word representations from different, complementary modalities, with very different coverages, but all including a small overlap. We assume that the different modalities contain complementary human text processing traces because they reflect different cognitive processes, which motivates us to combine these different sources of information. Our assumption is confirmed in the evaluation. The fact that we have very low coverage for some modalities, and the fact that we have an overlap between all our vocabularies, specifically motivates an approach, in which we use the intersection of word types to learn a projection from two or more of these modalities into a shared space. Obviously, we can also simply concatenate our representations, but because of the low coverage of some modalities and because co-projecting modalities has some regularization effect, we hypothesize that it is better to learn a projection into a shared space. This hypothesis is verified by the results in Section 8.6.

8.4.1 *Concatenating modalities*

The simplest way of combining the modalities is concatenating the corresponding vectors for each word. The different modalities have different dimensionalities, so we would need to perform dimensionality reduction to sum or average vectors, and the non-overlapping words don't allow for e.g. taking the outer product, so we simply concatenate the vectors instead. We use 0 for missing values.

8.4.2 *CCA*

Section 8.4.2 and Section 8.4.3 describe two different projection methods for projecting the representations in the different modalities into a shared space. We use the intersection of the lower-cased vocabulary for the alignment, i.e., as a supervision signal. For example, if the words *man*, *dog* and *speak* exist in both eigenword and keystroke data, from these 2 x 3 vectors, CCA estimate the transformation for the vectors for *house*, *cat* and *boy*, which (in this example) only exists in the keystroke data.

CCA, as originally proposed by Hotelling (1936), is a method of finding the optimum linear combination between two sets of variables, so the set of variables are transformed onto a projected space while the correlation is maximized. We use the implementation of Faruqui and Dyer (2014b) made for creating bilingual embeddings. We use modalities instead of languages. The size of the projected space is smaller than or equal to the original dimension.

We incrementally combine modalities and project them to new, shared spaces using the intersection of the lower-cased vocabulary. We add them by the order of word type count starting with the modality with most word types. For the first projection only, we reduce the size of the projected space. We set the ratio of the first projected space (only two modalities) to 0.6 based on POS induction results on development data using the setup described in Section 8.5.

8.4.3 *SVD and Inverted Softmax*

As an alternative to CCA, but closely related, we also use a projection method proposed and implemented by Smith et al. (2017), which uses (singular value decomposition and inverted softmax feature projection (SVD+IS)). This method uses a reference space, rather than projecting all modalities into a new space.

Smith et al. (2017) apply SVD+IS to obtain an orthogonal transformation matrix that maps the source language into the target language. In addition, in order to estimate their confidence on the predicted target, they use an inverted softmax function for determining the probability that a target word translates back into a source word.

Like for CCA, we incrementally project datasets onto each other starting with the most word-type rich modality. We use the highest dimensionality of any of our representations (88 dimensions).

8.5 EXPERIMENTS

This section presents our POS and syntactic chunk induction experiments. We present the datasets we used in our experiments, the sequence tagging architecture, based on second-order HMM, as well as the dictionary we used to constrain inference at training and test time.

8.5.1 *Data*

For unsupervised POS induction, we use Ontonotes 5.0 (Weischedel et al., 2013) for training, development and test. We set all hyper-parameters on the newswire (NW) domain, optimizing performance on the development set. Size of the development set is 154,146 tokens. We run individual experiments on each of the seven domains, with these hyper-parameters, reporting performance on the relevant

		Rules
DET	→	NP
VERB	→	VP
NOUN|PRONOUN|NUM	→	NP
.	→	O
ADJ	→	NP|ADJP
ADV	→	NP|VP|ADVP|AD
PRT	→	NP|PRT
CONJ	→	O|NP
ADP	→	PP|VP|SBAR

Table 8.2: Heuristics for expanding our POS dictionary to chunks

test set. The domains are broadcast conversation (BC), broadcast news (BN), magazines (MZ), newswire (NW), the Bible (PT), telephone conversations (TC), and weblogs (WB). We also train and test unsupervised POS induction on the CoNLL 2007 (Nivre et al., 2007) splits of the PTB (Marcus, Marcinkiewicz, and Santorini, 1993) using the hyper-parameter settings from Ontonotes. We mapped all POS labels to Google's coarse-grained, universal POS tagset (Petrov, Das, and McDonald, 2011). For model selection, we select based both on best results on Ontonotes NW development as well as PTB development sets.

For syntactic chunk induction, we use the bracketing data from PTB with the standard splits for syntactic chunking. We tune hyperparameters for chunking on the development set and select best models based on the development result.

8.5.2 *Model*

We used a modification of the implementation of a type-constrained, SHMM-ME from Li, Graça, and Taskar (2012). It is a second-order version of the first order maximum entropy HMM presented in (Berg-Kirkpatrick et al., 2010) with the important addition that it is constrained by a crowd-sourced tag dictionary (Wiktionary). This means that for all words in the Wiktionary, the model is only allowed to predict one of the tags listed for it in Wiktionary

The same model was used in Barrett et al. (2016) to improve unsupervised POS inducing using gaze data from the Dundee Corpus, and in Bingel, Barrett, and Søgaard (2016) to augment an unsupervised POS tagger with features from fMRI recordings.

The number of EM iterations used for inducing our taggers was tuned using eigenvector embeddings on the development data, considering values 1..50. POS performance peaked at iterations 30 and 31.

FEATURE SET	TA
No embeddings	60.32
Eigenwords	59.26
Best combined models	
CCA Dun_GECO_Pros	**63.33***†
SVD+IS GECO_Key_Pros	62.91*
Concat Eig_GECO_Key	61.16

Table 8.3: Chunk tagging accuracy. Best models from CCA, SVD+IS and concatenation. Model section on development set. * $p < .001$ Mcnemar mid-p test when comparing to no embeddings. † $p < .001$ Mcnemar mid-p test when comparing to Eigenwords.)

We use 30 in all our POS experiments. For syntactic chunking, we use 48 iterations, which led to the best performance on the PTB development data using only eigenword embeddings.

8.5.3 *Wiktionary*

The Wiktionary constrains the predicted tags in our model. The better the Wiktionary, the better the predictions.

For POS-tagging we used the same Wiktionary dump[4] that Li, Graça, and Taskar (2012) used in their original experiments. The Wiktionary dump associated word types with Google's universal parts-of-speech labels.

For chunking, Wiktionary does not provide direct information about the possible labels of words. We instead apply simple heuristics to relate POS information to syntactic chunking labels. Since we already know the relation between words and POS labels from Wiktionary, we can compute the transitive closure in order to obtain a dictionary relating words with syntactic chunking labels. We present the heuristics in Table 8.2.

Note that the rules are rather simple. We do not claim this is the best possible mapping. We are relying on these simple heuristics only to show that it is possible to learn syntactic chunkers in an unsupervised fashion by relying on a combination of features from different modalities and a standard, crowd-sourced dictionary.

4 https://code.google.com/archive/p/wikily-supervised-pos-tagger/

	Ontonotes								PTB
Feature set	bc	bn	mz	nw	pt	tc	wb	avg	
No embeddings	83.1	84.41	85.32	84.94	85.14	77.8	85.93	83.81	82.83
Eigenwords	83.16	84.68*	85.48	85.07	85.31	78.07	85.88	83.95	83.38*
Best Ontonotes NW models									
CCA Eig_Dun	**83.45***†	**84.99***	85.79*	**85.38***†	85.2	77.99	**86.38***†	84.17	**84.28***†
SVD+IS Dun_GECO_Key	83.24	84.76	**86.22***†	85.33*†	85.44	77.84	85.95	84.11	84.25*†
Concat Eig_Dun_GECO	83.39*†	84.78*	85.8*†	85.36*†	**85.45**	**78.38***	86.21†	**84.19**	83.91*†
Best PTB models									
CCA Eig_Dun	**83.45***†	**84.99***	85.79*	**85.38***†	85.2	77.99	**86.38***†	84.17	**84.28***†
SVD+IS Dun_Key	83.24	84.59	86.12*†	85.28*†	85.39	77.90	85.86	84.05	84.24*†
Concat Eig_Pros	83.22	84.54	85.67	85.01	84.98	77.98	85.97	83.91	84.22*†

Table 8.4: POS tagging accuracies for baselines and the model combinations that performed best on newswire development data (NW). Best performance per domain is boldfaced. *) $p < .001$ McNemar mid-p test when compared to the no embeddings condition for the corresponding test set. †) $p < .001$ McNemar mid-p test when compared to eigenwords for the corresponding test set.

8.6 RESULTS

All our POS tagging accuracies can be seen in Table 8.4. Our first observation is that human processing data helps unsupervised POS induction. In fact, the models augmented with processing data are *consistently* better than the baseline without vector representations, as well as better than only using distributional word embeddings.

Generally, CCA seems to find the best projection into a common space for system combinations. For PTB, the CCA-aligned model is the best and this result is significant ($p < .001$) when comparing both to no embeddings and eigenwords. For Ontonotes 5.0, CCA is better than the other projection methods in 4/7 domains, but when averaging, concatenation gets the higher result.

The standard embeddings are often part of the best combinations, but the human processing data contributes with important information; in 4/7 domains as well as on PTB data, we see a significantly better performance ($p < .001$) with a combination of modalities when comparing to eigenwords.

Aligning Dundee with eigenwords is the best POS model both according to the Ontonotes 5.0 NW development set and the PTB development set. Dundee is the most frequent modality in the six best POS induction models with five appearances. Eigenwords is second most frequent with four appearances.

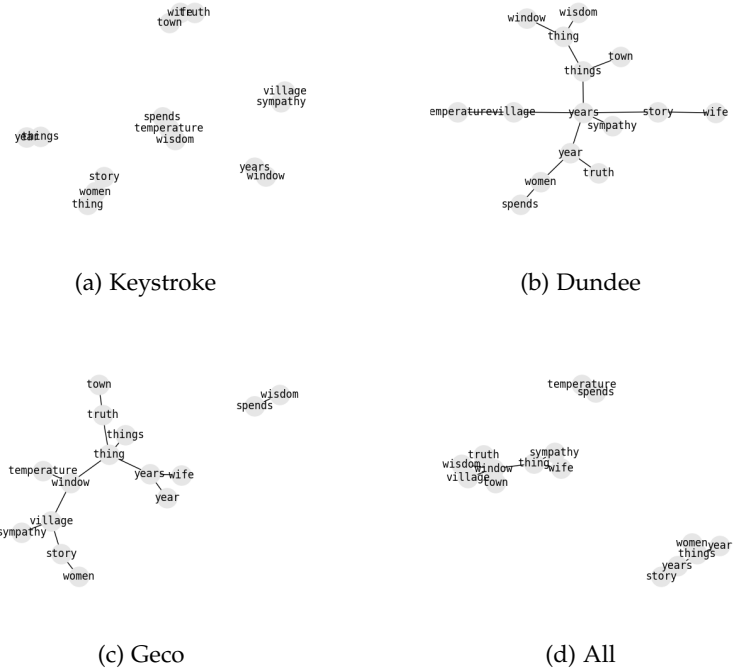
The syntactic chunking accuracies are in Table 8.3. Also here CCA is the better combination method. For chunking, all combined models are better than no embeddings and eigenwords. The improvement is significant compared to no embeddings except for concatenation ($p <$

	KEYSTROKE	DUNDEE	GECO
DUNDEE	16.84		
GECO	11.39	1.02	
CCA ALL	13.98	3.72	3.09

Table 8.5: Graph similarities in $[0, \infty)$, $0 =$ identical.

.001). For CCA, the result is significantly better than no embeddings and eigenwords.

For chunking, GECO data appears in all best models and is thus the most frequent modalities. Keystroke and prosody appears in two best models each.



(a) Keystroke

(b) Dundee

(c) Geco

(d) All

Figure 8.2: Nearest neighbor graphs for 15 frequent nouns.

8.7 ANALYSIS

8.7.1 *POS error analysis*

In this subsection we will explore what is learnt in the best model (CCA-alligned *eigen_dundee*) compared to the eigenwords baseline. We analyse the POS predictions on the seven Ontonotes development sets and compare the best model to the model using only eigenwords

	best model vs. baseline	Dev.
.	0.0	11.05
ADJ	10.13	7.22
ADP	8.40	11.51
ADV	11.22	5.62
CONJ	1.44	3.65
DET	2.28	10.31
NOUN	31.22	23.89
NUM	0.32	1.03
PRON	0.80	6.37
PRT	0.83	2.58
VERB	33.37	16.52
X	0.0	0.25

Table 8.6: Distribution of POS tags in the subset where the best model made correct predictions and the baseline made wrong predictions compared to the general development set distribution.

embeddings. The grand mean accuracy for the eigenwords baseline is 87.30. The best model is slightly better with 87.49.

There are 3120 words that are predicted correctly by the best model and wrong by the baseline model. We'll look into these specifically. Table 8.6 show the distribution of POS tags in this improved subset compared to the general distribution in the development sets. Punctuation is not improved, since we use a similar punctuation template for both models. For adjectives, adverbs, nouns, and verbs we see improvements compared to the overall distribution. Especially nouns and verbs, which are also the most numerous classes, are overrepresented in the improved subset.

6.52% of the words in the improved subset are not represented in the tag dictionary. In the general development set 3.27% of the words are not in the tag dictionary. This suggests that the best model does better for words that are not in the tag dictionary than the baseline model.

8.7.2 *What is in the vectors?*

NEAREST NEIGHBOR GRAPHS We include a detailed analysis of subgraphs of the nearest neighbor graphs in the embedding spaces of keystrokes, Dundee, GECO, and CCA projection of all modalities. Specifically, we consider the nearest neighbor graphs among the 15

most frequent unambigous nouns, according to Wiktionary.[5] See Figure 8.2 for plots of the nearest neighbor graphs. The prosody features containing less than 600 word types only contained 2 of the 15 nouns and is therefore not included in this analysis.

Projecting word representations into a shared space using linear methods assumes approximate isomorphism between the embedding spaces - or at least their nearest neighbor graphs. We use the VF2 algorithm (Cordella et al., 2001) to verify that the subgraphs are *not* isomorphic, but this can also be seen directly from Figure 8.2. Neither keystroke and gaze embeddings, nor the two different gaze-induced embeddings are isomorphic.

Since none of the modalities induce isomorphic nearest neighbor graphs, this does not tell us much about similarities between modalities. To quantify the similarity of non-isomorphic graphs, we use *eigenvector similarity* (Shigehalli and Shettar, 2011), which we calculate by computing the Laplacian eigenvalues for the nearest neighbors, and for each graph, find the smallest k such that the sum of the k largest eigenvalues is $<90\%$ of the eigenvalues. We then take the smallest k of the two, and use the sum of the squared differences between the largest k eigenvalues as our similarity metric.

Using this metric to quantify graph similarity, we see in Table 8.5 that, not surprisingly, the gaze graphs are the most similar. The projected space is more similar to the gaze spaces, but balances gaze and keystroke information. The GECO embeddings agree more with the keystrokes than the Dundee embeddings does.

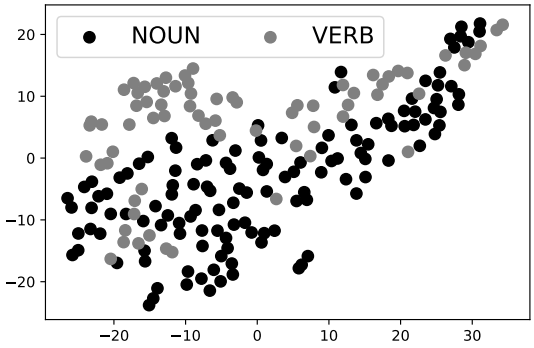
T-SNE PLOTS We take words that – according to the Wiktionary – can only have one tag and sort them by BNC frequency (Kilgarriff, 1995) in descending order. For these words and their POS tags we get the feature vector of the POS model yielding the highest result on both Ontonotes and PTB: CCA-projected eigenwords and Dundee features. For the first 200 occurrences of the frequency-sorted list, we reduce dimensionality using t-Distributed Stochastic Neighbor Embedding (t-SNE) (van der Maaten and Hinton, 2008) and plot the result. Figure 8.3 shows that 200 most frequent content words cluster with respect to their POS tag, somewhat distinguishing verbs from nouns and adjectives from adverbs in CCA space.

8.7.3 *How big a Wiktionary do we need?*

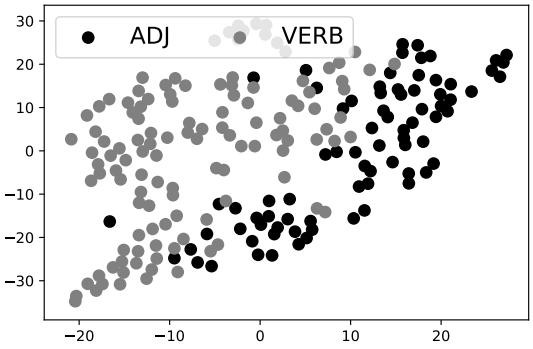
Our Wiktionary for English contains POS information for 72,817 word types. Word types have 6.2 possible POS categories on average meaning we have over 450.000 entries in our POS dictionary. For PTB, 70.0% of wordtypes of the test set are covered by the dictionary. For the

5 Wiktionary is a crowd-sourced, imperfect dictionary, and one of the "unambiguous nouns" is *spends*, which, we assume, you are more likely to encounter as a verb.

chunking data, 70.4% of wordtypes of the test set are covered by the dictionary. The English Wiktionary is thus much bigger than wiktionaries for low-resource language (Garrette and Baldridge, 2013). How big a dictionary is needed to achieve good performance, and can we get away with a smaller dictionary if we have processing data? This section explores the performance of the model as a function of the Wiktionary size.



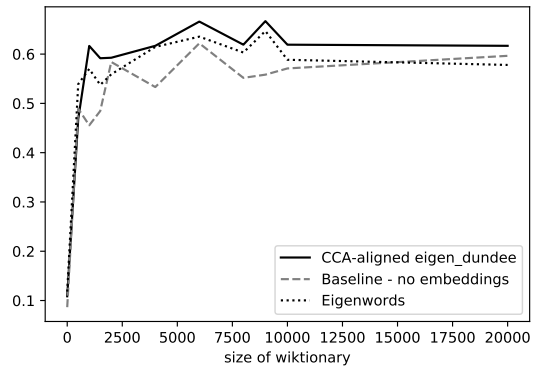
(a) NOUN and VERB



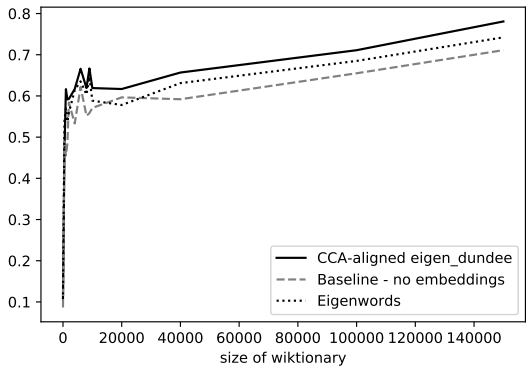
(b) ADJ and VERB

Figure 8.3: t-SNE plots of CCA-projected eigen dundee features for pairs of tags.

We sorted the Wiktionary by word frequency obtained from BNC (Kilgarriff, 1995) and increased the Wiktionary size for the best POS system starting with 0 (no dictionary). For each Wiktionary size, we compare with the baseline without access to processing data and eigenwords. The learning curve can be seen in Figure 8.4a and Figure 8.4b. We observe that having entries for the most frequent words is a lot better than having no dictionary, and that the difference between our best system and the baseline exists across all dictionary sizes. With 10,000 entries, all systems seems to reach a plateau.



(a) 0-20,000 entries



(b) 0-150,000 entries

Figure 8.4: Learning curve assuming Wiktionary entries for k most frequent words, comparing our best PoS induction system against our baseline. On Ontonotes WB development data, 30 training iterations.

8.8 DISCUSSION

GENRES AND DOMAINS When collecting our human language processing data, we did not control for genre. Our data sets span child-directed speech, free text composition, and skilled adults reading fiction and newspaper articles. The Dundee corpus (newspaper articles) matches the genre of at least some of the Ontonotes test set. Immonen and Mäkisalo (2010) found that for keystroke, genre does seem to have an effect on average pause length, be it sentence initial, word initial, clause initial or phrase initial. Texts organized linearly – e.g. reports and narratives – require less pausing than texts with a global approach, like expository, persuading and generalizing text. Our results show that human processing features transfer across genres, but within-genre data would probably be beneficial for results.

RICHER REPRESENTATIONS The type-level features we use do not take context into account, and the datasets we use, are too small to enrich our representations. Human processing data is more and more readily available, however. Eye trackers are probably built into the next generation of consumer hardware, and speech records and keystroke logs are recordable with existing technology.

8.9 CONCLUSION

We have shown how to improve unsupervised POS induction and syntactic chunking significantly using data reflecting human language processing. Our model, which is a second-order hidden Markov model, is the first to combine multidimensional, continuous features of eye movements, prosody and keystroke logs. We have shown that these features can be combined using projection techniques, even when they only partially overlap in word coverage. None of our models require access to these features at test time. We experimented with all combinations of modalities, and our results indicate that eye tracking is useful for both chunking and POS induction. Finally, we have shown that the potential impact of human processing data also applies in a low-resource setting, i.e., when available tag dictionaries are small.

ACKNOWLEDGEMENTS

Thanks to Desmond Elliott for valuable ideas. This research was partially funded by the ERC Starting Grant LOWLANDS No. 313695, as well as by Trygfonden.

Part IV

SEQUENCE CLASSIFICATION AND MODELLING REAL-WORLD DATA

9

SEQUENCE CLASSIFICATION WITH HUMAN ATTENTION

ABSTRACT

Learning attention functions requires large volumes of data, but many NLP tasks simulate human behavior, and in this paper, we show that human attention really does provide a good inductive bias on many attention functions in NLP. Specifically, we use estimated human attention derived from eye-tracking corpora to regularize attention functions in recurrent neural networks. We show substantial improvements across a range of tasks, including sentiment analysis, grammatical error detection, and detection of abusive language.

9.1 INTRODUCTION

When humans read a text, they do not attend to *all* its words (Carpenter and Just, 1983; Rayner and Duffy, 1988). For example, humans are likely to omit many function words and other words that are predictable in context and focus on less predictable content words. Moreover, when they fixate on a word, the duration of that fixation depends on a number of linguistic factors (Clifton, Staub, and Rayner, 2007; Demberg and Keller, 2008).

Since learning good attention functions for recurrent neural networks requires large volumes of data (Britz, Guan, and Luong, 2017; Zoph et al., 2016), and errors in attention are known to propagate to classification decisions (Alkhouli et al., 2016), we explore the idea of using human attention, as estimated from eye-tracking corpora, as an inductive bias on such attention functions. Penalizing attention functions for departing from human attention may enable us to learn better attention functions when data is limited.

Eye-trackers provide millisecond-accurate records on where humans look when they are reading, and they are becoming cheaper and more easily available by the day (San Agustin et al., 2009). In this paper, we use publicly available eye-tracking corpora, i.e., texts augmented with eye-tracking measures such as fixation duration times, and large eye-tracking corpora have appeared increasingly over the past years. Some studies suggest that the relevance of text can be inferred from the gaze pattern of the reader (Salojärvi et al., 2003) – even on word-level (Loboda, Brusilovsky, and Brunstein, 2011).

CONTRIBUTIONS We present a recurrent neural architecture with attention for sequence classification tasks. The architecture jointly learns its parameters and an attention function, but can alternate between supervision signals from labeled sequences (with no explicit supervision of the attention function) and from attention trajectories. This enables us to use per-word fixation durations from eye-tracking corpora to regularize attention functions for sequence classification tasks. We show such regularization leads to significant improvements across a range of tasks, including sentiment analysis, detection of abusive language, and grammatical error detection. Our implementation is made available at `https://github.com/coastalcph/Sequence_classification_with_human_attention`.

9.2 METHOD

We present a recurrent neural architecture that jointly learns the recurrent parameters and the attention function, but can alternate between supervision signals from labeled sequences and from attention trajectories in eye-tracking corpora. The input will be a set of labeled sequences (sentences paired with discrete category labels) and a set of sequences, in which each token is associated with a scalar value representing the attention human readers devoted to this token on average.

The two input datasets, i.e., the target task training data of sentences paired with discrete categories, and the eye-tracking corpus, need not (and will not in our experiments) overlap in any way. Our experimental protocol, in other words, does not require in-task eye-tracking recordings, but simply leverages information from existing, available corpora.

Behind our approach lies the simple observation that we can correlate the token-level attention devoted by a recurrent neural network, even if trained on sentence-level signals, with any measure defined at the token level. In other words, we can compare the attention devoted by a recurrent neural network to various measures, including token-level annotation (Rei and Søgaard, 2018) and eye-tracking measures. The latter is particularly interesting as it is typically considered a measurement of *human* attention.

We go beyond this: Not only can we compare machine attention with human attention, we can also constrain or inform machine attention by human attention in various ways. In this paper, we explore this idea, proposing a particular architecture and training method that, in effect, uses human attention to *regularize* machine attention.

Our training method is similar to a standard approach to training multi-task architectures (Bingel and Søgaard, 2017; Dong et al., 2015; Søgaard and Goldberg, 2016), sometimes referred to as the *alternating* training approach (Luong et al., 2016): We randomly select a data

point from our training data or the eye-tracking corpus with some (potentially equal) probability. If the data point is sampled from our training data, we predict a discrete category and use the computed loss to update our parameters. If the data point is sampled from the eye-tracking corpus, we still run the recurrent network to produce a category, but this time we only monitor the attention weights assigned to the input tokens. We then compute the minimum squared error between the normalized eye-tracking measure and the normalized attention score. In other words, in multi-task learning, we optimize each task for a fixed number of parameter updates (or mini-batches) before switching to the next task (Dong et al., 2015); in our case, we optimize for a target task (for a fixed number of updates), then improve our attention function based on human attention (for a fixed number of updates), then return to optimizing for the target task and continue iterating.

9.2.1 *Model*

Our architecture is a bi-LSTM (Hochreiter and Schmidhuber, 1997) that encodes word representations x_i into forward and backward representations, and into combined hidden states h_i (of slightly lower dimensionality) at every timestep. In fact, our model is a hierarchical model whose word representations are concatenations of the output of character-level LSTMs and word embeddings, following Plank, Goldberg, and Søgaard (2016), but we ignore the character-level part of our architecture in the equations below:

$$\overrightarrow{h_i} = LSTM(x_i, \overrightarrow{h_{i-1}}) \tag{9.1}$$

$$\overleftarrow{h_i} = LSTM(x_i, \overleftarrow{h_{i+1}}) \tag{9.2}$$

$$\widetilde{h_i} = [\overrightarrow{h_i}; \overleftarrow{h_i}] \tag{9.3}$$

$$h_i = \tanh(W_h \widetilde{h_i} + b_h) \tag{9.4}$$

The final (reduced) hidden state is sometimes used as a sentence representation s, but we instead use attention to compute s by multiplying dynamically predicted attention weights with the hidden states for each time step. The final sentence predictions y are then computed by passing s through two more hidden layers:

$$s = \sum_i \widetilde{a_i} h_i \tag{9.5}$$

$$y = \sigma(W_y \tanh(W_{\tilde{y}} s + b_{\tilde{y}}) + b_y) \tag{9.6}$$

From the hidden states, we directly predict token-level raw attention scores a_i:

$$e_i = \tanh(W_e h_i + b_e) \tag{9.7}$$

$$a_i = W_a e_i + b_a \tag{9.8}$$

We normalize these predictions to attention weights $\widetilde{a_i}$:

$$\widetilde{a_i} = \frac{a_i}{\sum_k a_k} \tag{9.9}$$

Our model thus combines two distinct objectives: one at the sentence level and one at the token level. The sentence-level objective is to minimize the squared error between output activations and true sentence labels $\widehat{y}$.

$$L_{sent} = \sum_j (y^{(j)} - \widehat{y}^{(j)})^2 \tag{9.10}$$

The token-level objective, similarly, is to minimize the squared error for the attention not aligning with our human attention metric.

$$L_{tok} = \sum_j \sum_t (a^{(j)(t)} - \widehat{a}^{(j)(t)})^2 \tag{9.11}$$

These are finally combined to a weighted sum, using λ (between 0 and 1) to trade off loss functions at the sentence and token levels.

$$L = L_{sent} + \lambda L_{tok} \tag{9.12}$$

Note again that our architecture does not require the target task data to come with eye-tracking information. We instead learn jointly to predict sentence categories and to attend to the tokens humans tend to focus on for longer. This requires a training schedule that determines when to optimize for the sentence-level classification objective, and when to optimize the machine attention at the token level. We therefore define an epoch to comprise a fixed number of batches, and sample every batch of training examples either from the target task data or from the eye-tracking corpus, as determined by a coin flip, the bias of which is tuned as a hyperparameter. Specifically, we define an epoch to consist of n batches, where n is the number of training sentences in the target task data divided by the batch size. This coin is potentially weighted, with data being drawn from the auxiliary task with some probability or a decreasing probability of $\frac{1}{E+1}$, where E is the current epoch; see Section 9.4 for hyper-parameters.

9.3 DATA

As mentioned in the above, our architecture requires no overlap between the eye-tracking corpus and the training data for the target task. We therefore rely on publicly available eye-tracking corpora. For sentiment analysis, grammatical error detection, and hate speech detection, we use publicly available research datasets that have been used previously in the literature. All datasets were lower-cased.

Task	Trains set		Dev. set		Test set	
	domain	*n* sent	domain	*n* sent	domain	*n* sent
Sentiment	SemEval Twitter	7177	SemEval Twitter	1,205	SemEval Twitter	2,870
Sentiment					SemEval SMS	2,094
Gram. err.	FCE	28,731	FCE	2,222	FCE	2,720
Hatespeech	Waseem (2016)	5,529	Waseem (2016)	690	Waseem (2016)	690
Hatespeech	Waseem and Hovy (2016)	11,225	Waseem and Hovy (2016)	1403	Waseem and Hovy (2016)	1,403

Table 9.1: Overview over the tasks and datasets used.

9.3.1 *Eye-tracking corpora*

For our experiments, we concatenate two publicly available eye-tracking corpora, the Dundee Corpus (Kennedy, Hill, and Pynte, 2003) and the reading parts of the ZuCo Corpus (Hollenstein et al., 2018), described below. Both corpora contain eye-tracking measurements from several subjects reading the same text. For every token, we compute the mean duration of all fixations to this token as our measure of human attention, following previous work (Barrett et al., 2016; Gonzalez-Garduno and Søgaard, 2018).

DUNDEE The English part of the Dundee corpus (Kennedy, Hill, and Pynte, 2003) comprises 2368 sentences and more than 50,000 tokens. The texts were read by ten skilled, adult, native speakers. The texts are 20 newspaper articles from *The Independent*. The reading was self-paced and as close to natural, contextualized reading as possible for a laboratory data collection. The apparatus was a Dr Bouis Oculometer Eyetracker with a 1000 Hz monocular (right) sampling. At most five lines were shown per screen while subjects were reading.

ZUCO The ZuCo corpus (Hollenstein et al., 2018) is a combined eye-tracking and EEG dataset. It contains approximately 1,000 individual English sentences read by 12 adult, native speakers. Eye movements were recorded with the infrared video-based eye tracker *EyeLink 1000 Plus* at a sampling rate of 500 Hz. The sentences were presented at the same position on the screen, one at a time. Longer sentences spanned multiple lines. The subjects used a control pad to switch to the next sentence and to answer the control questions, which allowed for natural reading speed. The corpus contains both natural reading and

reading in a task-solving context. For compatibility with the Dundee corpus, we only use the subset of the data, where humans were encouraged to read more naturally. This subset contains 700 sentences. This part of the Zuco corpus contains positive, negative or neutral sentences from the Stanford Sentiment Treebank (Socher et al., 2013) for passive reading, to analyze the elicitation of emotions and opinions during reading. As a control condition, the subjects sometimes had to rate the quality of the described movies; in approximately 10% of the cases. The Zuco corpus also contains instances where subjects were presented with Wikipedia sentences that contained semantic relations such as *employer*, *award* and *job_title* (Culotta, McCallum, and Betz, 2006). The control condition for this tasks consisted of multiple-choice questions about the content of the previous sentence; again, approximately 10% of all sentences were followed by a question.

PREPROCESSING OF EYE-TRACKING DATA Mean fixation duration (MEAN FIX DUR) is extracted from the Dundee Corpus. For Zuco, we divide total reading time per word token with the number of fixations to obtain mean fixation duration. The mean fixation duration is selected empirically among gaze duration (sum of all fixations in the first pass reading of the a word) and total fixation duration, and n fixations. Then we average these numbers for all readers of the corpus to get a more robust average processing time. Eye-tracking is known to correlate with word frequency (Rayner and Duffy, 1988). We include a frequency baseline on the eye tracking text, BNC INV FREQ. The word frequencies comes from the British National Corpus (BNC) frequency lists (Kilgarriff, 1995). We use log-transformed frequency per million. Before normalizing, we take the additive inverse of the frequency, such that rare words get a high value, making it comparable to gaze.

MEAN FIX DUR and BNC INV FREQ are min-max-normalized to a value in the range 0-1. MEAN FIX DUR is normalized separately for the two eye tracking corpora. We expect the experimental bias – especially the fact that ZuCo contains reading of isolated sentences and Dundee contains longer texts – to influence the reading and therefore separate normalization should preserve the signal within each corpus better.

9.3.2 *Sentiment classification*

Table 9.1 presents an overview of all train, development and test sets used in this paper.

Our first task is sentence-level sentiment classification. We note that many sentiment analysis datasets contain document-level labels or include more fine-grained annotation of text spans, say phrases or words. For compatibility with our other tasks, we focus on sentence-

level sentiment analysis. We use the SemEval-2013 Twitter dataset (Rosenthal et al., 2015; Wilson et al., 2013) for training and development. For test, we use a same-domain test set, the SemEval-2013 Twitter test set (SEMEVAL TWITTER POS | NEG), and an out-of-domain test set, SemEval-2013 SMS test set (SEMEVAL SMS POS | NEG). The SemEval-2013 sentiment classification task was a three-way classification task with positive, negative and neutral classes. We reduce the task to binary tasks detecting negative sentences vs. non-negative and vice versa. Therefore the dataset size is the same for POS and NEG experiments.

9.3.3 *Grammatical error detection*

Our second task is grammatical error detection. We use the First Certificate in English error detection dataset (FCE) (Yannakoudakis, Briscoe, and Medlock, 2011). This dataset contains essays written by English learners during language examinations, where any grammatical errors have been manually annotated by experts. Rei and Yannakoudakis (2016) converted the dataset for a sequence labeling task and we use their splits for training, development and testing. Similarly to Rei and Søgaard (2018), we perform sentence-level binary classification of sentences that need some editing vs. grammatically correct sentences. We do not use the token-level labels for training our model.

9.3.4 *Hate speech detection*

Our third and final task is detection of abusive language; or more specifically, hate speech detection. We use the datasets of Waseem (2016) and Waseem and Hovy (2016). The former contains 6,909 tweets; the latter 14,031 tweets. They are manually annotated for sexism and racism. In this study, sexism and racism are conflated into one category in both datasets. Both datasets are split in train, development and test splits consisting of 80%, 10% and 10% of the tweets respectively.

9.4 EXPERIMENTS

MODELS In our experiments, we compare three models: (a) a baseline model with automatically learned attention, (b) our model with an attention function regularized by information about human attention, and finally, (c) a second baseline using frequency information as a proxy for human attention and using the same regularization scheme as in our human attention model.

	BL			BNC INV FREQ			MEAN FIX DUR		
TASK	P	R	F_1	P	R	$F1$	P	R	F_1
SEMEVAL SMS NEG	43.55	45.41	43.77	45.82	48.65	45.24	47.15	46.98	**45.77**
SEMEVAL SMS POS	65.79	50.81	57.08	65.92	51.04	57.45	65.46	52.95	**58.50**
SEMEVAL TWITTER NEG	57.39	26.87	35.70	62.50	28.66	37.78	60.52	30.67	**40.23**
SEMEVAL TWITTER POS	77.96	53.88	63.63	79.66	54.66	64.78	78.77	55.35	**64.96**
FCE	79.01	89.33	83.84	79.18	89.26	83.89	79.03	90.28	**84.28**
WASEEM (2016)	76.42	62.07	68.29	77.20	61.71	68.54	77.20	63.06	**69.30**
WASEEM AND HOVY (2016)	76.23	72.23	74.16	76.33	74.70	75.48	76.95	74.43	**75.61**
MEAN	68.05	57.23	60.92	69.52	58.38	61.88	69.30	59.10	**62.67**

Table 9.2: Sentence classification results. P(recision), R(ecall) and F_1. Averages over 10 random seeds. Best average F_1 score per task is shown in bold.

HYPERPARAMETERS Basic hyper-parameters such as number of hidden layers, layer size, and activation functions were following the settings of Rei and Søgaard (2018). The dimensionality of our word embedding layer was set to size 300, and we use publicly available pre-trained Glove word embeddings (Pennington, Socher, and Manning, 2014) that we fine-tune during training. The dimensionality of the character embedding layer was set to 100. The recurrent layers in the character-level component have dimensionality 100; the word-level recurrent layers dimensionality 300. The dimensionality of our feed-forward layer, leading to reduced combined representations h_i, is 200, and the attention layer has dimensionality 100.

Three hyper-parameters, however, we tune for each architecture and for each task, by measuring sentence-level F_1-scores on the development sets. These are: (a) learning rate, (b) λ in Equation 9.12, i.e., controlling the relative importance of the attention regularization, and (c) the probability of sampling data from the eye-tracking corpus during training.

For all tasks and all conditions (baseline, frequency-informed baseline, and our human attention model), we perform a grid search over learning rates [.01 .1 1.], L_{att} weight λ values [.2 .4 .6 .8 1.], and probability of sampling from the eye-tracking corpus [.125 .25 .5 1., decreasing] – where *decreasing* means that the probability of sampling from the eye-tracking corpus initially is 0.5, but drops linearly for each epoch ($\frac{1}{E+1}$; see Section 9.2.1). We apply the models with the best average F_1 scores over three random seeds on the validation data, to our test sets.

INITIALIZATION Our models are randomly initialized. This leads to some variance in performance across different runs. We therefore report averages over 10 runs in our experiments below.

9.5 RESULTS

Our performance metric across all our experiments is the sentence-level F_1 score. We report precision, recall and F_1 scores for all tasks in Table 9.2.

Our main finding is that our human attention model, based on regularization from mean fixation durations in publicly available eye-tracking corpora, consistently outperforms the recurrent architecture with learned attention functions. The improvements over both baseline and BNC frequency are significant ($p < 0.01$) using bootstrapping (Calmettes, Drummond, and Vowler, 2012) over all tasks, with one seed. The mean error reduction over the baseline is 4.5%.

Unsurprisingly, knowing that human attention helps guide our recurrent architecture, the frequency-informed baseline is also better than the non-informed baseline across the board, but the human attention model is still significantly better across all tasks ($p < 0.01$). For all tasks except negative sentiment, we note that generally, most of the improvements over the learned attention baseline for the gaze-informed models, are due to improvements in recall. Precision is not worse, but we do not see any larger improvements on precision either. For the negative SEMEVAL tasks, we also see larger improvements for precision.

The observation that improvements are primarily due to increased recall, aligns well with the hypothesis that human attention serves as an efficient regularization, preventing overfitting to surface statistical regularities that can lead the network to rely on features that are not there at test time (Globerson and Roweis, 2006), at the expense of target class precision.

9.6 ANALYSIS

We illustrate the differences between our baseline models and the model with gaze-informed attention by the attention weights of an example sentence. Though it is a single, cherry-picked example, it is representative of the general trends we observe in the data, when manually inspecting attention patterns. Table 9.3 presents a coarse visualization of the attention weights of six different models, namely our baseline architecture and the architecture with gaze-informed attention, trained on three different tasks: hate speech detection, negative sentiment classification, and error detection. The sentence is a positive hate speech example from the Waseem and Hovy (2016) development set. The words with more attention than the sentence average are bold-faced.

First note that the baseline models only attend to one or two coherent text parts. This pattern was very consistent across all the sentences

FCE		SemEval Twitter NEG		Waseem and Hovy (2016)	
BL	MFD	BL	MFD	BL	MFD
@CharlesClassiqk:	**@CharlesClassiqk:**	**@CharlesClassiqk:**	**@CharlesClaqqqqqqqssiqk:**	@CharlesClassiqk:	@CharlesClassiqk:
sorry	**sorry**	**sorry**	**sorry**	sorry	sorry
I'm	I'm	**I'm**	**I'm**	I'm	I'm
not	not	**not**	**not**	not	not
sexist	**sexist**	**sexist**	**sexist**	sexist	sexist
BUT	BUT	BUT	BUT	BUT	**BUT**
there	**there**	there	there	there	**there**
is	is	is	is	is	is
a	a	a	a	a	a
double	**double**	double	**double**	double	**double**
standards	**standards**	standards	**standards**	**standards**	**standards**
there's	**there's**	there's	there's	**there's**	**there's**
certain	**certain**	certain	**certain**	**certain**	**certain**
rules	rules	rules	**rules**	**rules**	**rules**
for	for	for	for	**for**	**for**
dudes	dudes	dudes	dudes	**dudes**	**dudes**
and	and	and	and	**and**	and
there's	**there's**	there's	there's	**there's**	**there's**
certain	**certain**	certain	**certain**	**certain**	**certain**
rules	**rules**	rules	**rules**	**rules**	**rules**
for	for	for	for	**for**	for
fem...	**fem...**	fem...	fem...	**fem...**	fem...

Table 9.3: One sentence marked as containing sexism from Waseem and Hovy (2016) development set. Using trained baseline (BL) and gaze model (MFD) for three tasks: error detection, sentiment classification, and hate speech detection. Words with more attention than sentence average are boldfaced.

we examined. This pattern was not observed with gaze-informed attention.

Our second observation is that the baseline models are more likely to attend to stop words than gaze-informed attention. This suggests that gaze-informed attention has learned to simulate human attention to some degree. We also see many differences between the jointly learned task-specific, gaze-informed attention functions.

The gaze-informed hate speech classifier, for example, places considerable attention *BUT*, which in this case is a passive-aggressive hate speech indicator. It also gives weight to *double standards* and *certain rules*.

The gaze-informed sentiment classifier, on the other hand, focuses more on *sorry I am not sexist* which, in isolation, reads like an apologetic disclaimer. This model also gives weight to *double standards* and *certain rules*

The gaze-informed grammatical error detection model gives attention to *standards*, which is ungrammatical, because of the morphological number disagreement with its determiner *a*; it also gives attention to *certain rules*, which is disagreeing, again in number, with *there's*. It also gives attention to the non-word *fem*.

Overall, this, in combination with our results in Table 9.3, suggests that the regularization effect from human attention enables our architecture to learn to better attend to the most relevant aspects of sentences for the target tasks. In other words, human attention provides the inductive bias that makes learning possible.

9.7 DISCUSSION AND RELATED WORK

GAZE IN NLP It has previously been shown that several NLP tasks benefit from gaze information, including part-of-speech tagging (Barrett and Søgaard, 2015b; Barrett et al., 2016), prediction of MWEs (Rohanian et al., 2017) and sentiment analysis (Mishra et al., 2016b).

Gaze information and other measures from psycholinguistics have been used in different ways in NLP. Some authors have used discretized, single features (Klerke, Goldberg, and Søgaard, 2016; Pate and Goldwater, 2011, 2013; Plank, 2016a), whereas others have used multidimensional, continuous values (Barrett et al., 2016; Bingel, Barrett, and Søgaard, 2016). We follow Gonzalez-Garduno and Søgaard (2018) in using a single, continuous feature. We did not experiment with other representations, however. Specifically, we only considered the signal from token-level, normalized mean fixation durations.

Fixation duration is a feature that carries an enormous amount of information about the text and the language understanding process. Carpenter and Just (1983) show that readers are more likely to fixate on open-class words that are not predictable from context, and Kliegl et al. (2004) show that a higher cognitive load results in longer fixation durations. Fixations before skipped words are shorter before short or high-frequency words and longer before long or low-frequency words in comparison with control fixations (Kliegl and Engbert, 2005). Many of these findings suggest correlations with syntactic information, and many authors have confirmed that gaze information is useful to discriminate between syntactic phenomena (Barrett and Søgaard, 2015a,b; Demberg and Keller, 2008).

Gaze data has also been used in the context of sentiment analysis before (Mishra, Dey, and Bhattacharyya, 2017; Mishra et al., 2016b). Mishra et al. (2016b) augmented a sentiment analysis system with eye-tracking features, including first fixation durations and fixation counts. They show that fixations not only have an impact in detecting sentiment, but also improve sarcasm detection. They train a convolutional neural network that learns features from both gaze and text and uses them to classify the input text (Mishra, Dey, and Bhattacharyya, 2017). On a related note, Raudonis et al. (2013) developed a emotion recognition system from visual stimulus (not text) and showed that features such as pupil size and motion speed are relevant to accurately detect emotions from eye-tracking data. Wang, Zhang, and Zong (2017) use variables shown to correlate with human attention, e.g. surprisal, to guide the attention for sentence representations.

Gaze has also been used in the context of grammaticality (Klerke, Alonso, and Søgaard, 2015; Klerke et al., 2015), as well as in readability assessment (Gonzalez-Garduno and Søgaard, 2018).

Gaze has either been used as features (Barrett, Keller, and Søgaard, 2016; Barrett and Søgaard, 2015a) or as a direct supervision signal in

multi-task learning scenarios (Gonzalez-Garduno and Søgaard, 2018; Klerke, Goldberg, and Søgaard, 2016). We are, to the best of our knowledge, the first to use gaze to inform attention functions in RNNs.

HUMAN-INSPIRED ATTENTION FUNCTIONS Ibraheem, Altieri, and DeNero (2017), however, uses optimal attention to simulate human attention in an interactive machine translation scenario, and Britz, Guan, and Luong (2017) limit attention to a local context, inspired by findings in studies of human reading. Rei and Søgaard (2018) use auxiliary data to regularize attention functions in RNNs; not from psycholinguistics data, but using small amounts of task-specific, token-level annotations. While their motivation is very different from ours, technically our models are very related. In a different context, Das et al. (2017) investigated whether humans attend to the same regions as neural networks solving visual question answering problems. Lindsey (2017) also used human-inspired, unsupervised attention in a computer vision context.

OTHER WORK ON MULTI-PURPOSE ATTENTION FUNCTIONS While our work is the first to use gaze data to guide attention in a recurrent architectures, there has recently been some work on sharing attention functions across tasks. Firat, Cho, and Bengio (2016), for example, share attention functions between languages in the context of multi-way neural machine translation.

SENTIMENT ANALYSIS While sentiment analysis is most often considered a supervised learning problem, several authors have leveraged other signals than annotated data to learn sentiment analysis models that generalize better. Felbo et al. (2017), for example, use emoji prediction to pretrain their sentiment analysis models. Mishra et al. (2018) use several auxiliary tasks, including gaze prediction, for document-level sentiment analysis. There is a lot of previous work, also, leveraging information across different sentiment analysis datasets, e.g., Liu, Qiu, and Huang (2016).

ERROR DETECTION In grammatical error detection, Rei (2017) used an unsupervised auxiliary language modeling task, which is similar in spirit to our second baseline, using frequency information as auxiliary data. Rei and Yannakoudakis (2017) go beyond this and evaluate the usefulness of many auxiliary tasks, primarily syntactic ones. They also use frequency information as an auxiliary task.

HATE SPEECH DETECTION In hate speech detection, many signals beyond the text are often leveraged (see Schmidt and Wiegand (2017) for an overview of the literature). Interestingly, many authors have used signals from sentiment analysis, e.g., Gitari et al. (2015), moti-

vated by the correlation between hate speech and negative sentiment. This correlation may also explain why we see the biggest improvements with gaze-informed attention on those two tasks.

HUMAN INDUCTIVE BIAS Finally, our work relates to other work on providing better inductive biases for learning human-related tasks by observing humans (Tamuz et al., 2011; Wilson et al., 2015). We believe this is a truly exciting line of research that can help us push research horizons in many ways.

9.8 CONCLUSION

We have shown that human attention provides a useful inductive bias on machine attention in RNNs for sequence classification problems. We present an architecture that enables us to leverage human attention signals from general, publicly available eye-tracking corpora, to induce better, more robust task-specific NLP models. We evaluate our architecture and show improvements across three NLP tasks, namely sentiment analysis, grammatical error detection, and detection of abusive language. We observe that not only does human attention help models distribute their attention in a generally useful way; human attention also seems to act like a regulariser providing more robust performance across domains, and it enables better learning of task-specific attention functions through joint learning.

10

PREDICTING MISREADINGS FROM GAZE IN CHILDREN WITH READING DIFFICULTIES

ABSTRACT

We present the first work on predicting reading mistakes in children with reading difficulties based on eye-tracking data from real-world reading teaching. Our approach employs several linguistic and gaze-based features to inform an ensemble of different classifiers, including multi-task learning models that let us transfer knowledge about individual readers to attain better predictions. Notably, the data we use in this work stems from noisy readings *in the wild*, outside of controlled lab conditions. Our experiments show that despite the noise and despite the small fraction of misreadings, gaze data improves the performance more than any other feature group and our models achieve good performance. We further show that gaze patterns for misread words do not fully generalize across readers, but that we can transfer some knowledge between readers using multi-task learning at least in some cases. Applications of our models include partial automation of reading assessment as well as personalized text simplification.

10.1 INTRODUCTION

Reading disabilities are impairments affecting individuals' access to written sources, with downstream effects such as low self-confidence in the classroom and limited access to higher education. Dyslexia, for instance, while being highly prevalent with estimates reaching up to 17.5% of the entire population of the U.S. (Interagency Committee on Learning Disabilities, 1987), often goes undiagnosed, such that unattributed weaknesses in reading comprehension further intimidate affected persons. Due to these severe and broad-ranging impacts of reading difficulties, many governments have implemented early screening tests for dyslexia and other reading difficulties and provide special training and assistance for struggling readers throughout the educational system and into adulthood.

In Denmark, for example, such programs provide children with specialist training through focused multi-week reading courses in one-on-one or small group settings. Still, the specialized teachers can only attend to one student at a time when closely monitoring their reading, and the quality of any analysis is strictly limited by the

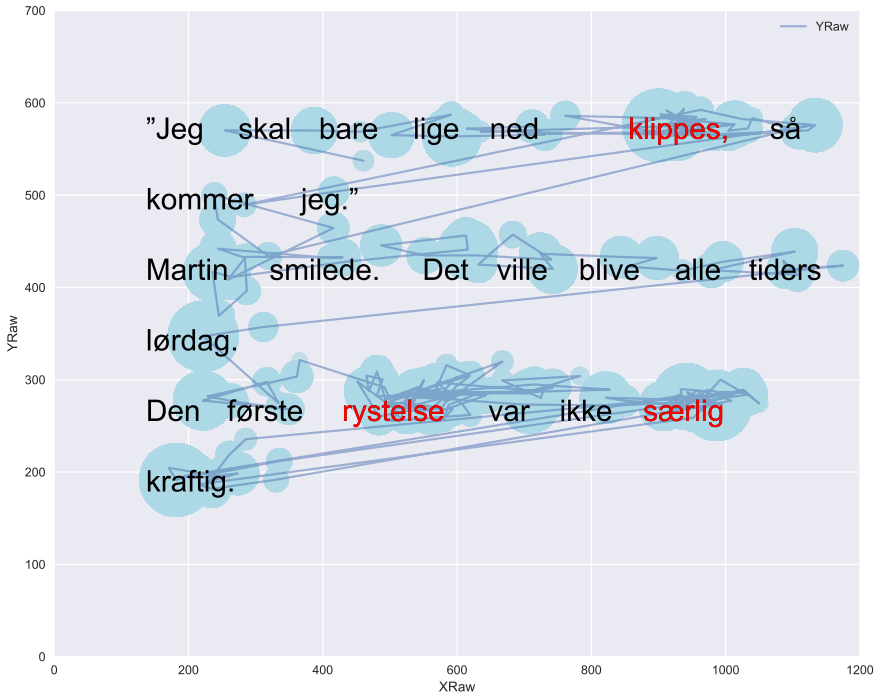


Figure 10.1: Scanpath and fixations (blue circles) when reading a sentence. This particularly clear example from our dataset shows extended processing time for misread words (marked in red).

human observer's processing "bandwidth" while attending the live reading.

As a possible mitigation, advances in eye-tracking technology – in particular the increased availability of eye trackers – have made it possible to reliably record children's gaze during reading, both allowing teachers to attend to their students' reading post-hoc as well as providing additional insight into reading strategies based on gaze, including the development of these strategies over time. For the teacher to track and keep records of reading mistakes (henceforth referred to as *misreadings*), however, the students are still required to read out loud, and the teacher has to review the entire reading and annotate for misreadings.

In this work, we investigate to what extent we can predict misreadings from gaze patterns for individual words. While the aim is not to fully automate reading reviews, being able to successfully predict misreadings from gaze data can be part of a semi-automatic system for reading quality assessment and increase teacher efficiency by pointing out potential misreadings for closer review.

Another motivation for this work comes from text simplification, in particular from the observation that individuals' highly specific reading strengths and weaknesses require text simplification models to be customized to specific users in order to unfold their full potential and truly be helpful. Predicting misreadings in concrete reading sce-

narios and based on individual gaze patterns can be used as a first step in the typical lexical simplification pipeline (Shardlow, 2014).[1] This task, known as complex word identification, has received a considerable amount of attention in the literature, but has exclusively been approached in a user-agnostic fashion.

The data used in this study are gaze recordings of children with reading difficulties, reading Danish texts assigned by their reading teacher as part of their reading intervention. The recordings stem from EyeJustRead, an eye-tracking based software used in special reading intervention in Danish schools.[2] In Section 10.3, we discuss further aspects of the treatment of gaze data in general and the collection of the data used in this study in particular.

While the difficulty of processing a word is undoubtedly reflected in the fixation time on that word (Rayner et al., 1989), many other factors affect fixation durations, the most prominent being word length and word frequency, but also predictability and relative position in sentence have strong effects–see Figure 10.1 for a particularly clear example from our dataset. Notably, almost all analyses of eye-tracking reading data use data collected in research laboratories, where these–otherwise confounding–factors can be controlled for. We show that we can perform reasonable misreading detection on real-world eye tracking data, including a limited number of textual features to control for these factors.

CONTRIBUTIONS

A. We present the first work on the automatic detection of misreadings based on gaze patterns of children with reading difficulties.

B. This is, to the best of our knowledge, the first attempt at modeling noisy, real-world eye-tracking data from readers.

C. We also present, to the best of our knowledge, the first published results using a multi-task learning setup to transfer knowledge between individual readers for personalized, complex word identification.

10.2 RELATED WORK

Our work is a special case of complex word identification, a task that has recently received a significant amount of interest, including two shared tasks (Paetzold and Specia, 2016; Yimam et al., 2018). The

1 While today it may hardly sound plausible to equip each laptop with an eye-tracker in order to track people's reading, further technological advances may well make this possible in the future. Recent development in eye-tracking technology has taken it from expensive research equipment to a gaming interface with a price point as low as $100.

2 http://www.eyejustread.com

most successful approaches to these tasks had in common that they employed ensembles of classifiers that learned from a number of semantic and psycholinguistic features. Note however, that these previous approaches to complex word identification aimed at developing generic models that took no account of any specifics of a certain user.

Children's eye movements during reading are not as well-studied as adults', and previous studies typically analyze data collected in experiments designed for research. The overall established observations with regards to reading development are: older children have shorter fixation durations, fewer fixations and fewer regressions. They have a higher skipping probability and also higher saccade amplitude. See Blythe and Joseph (2011) for a review. It is not conclusive whether these variations follow chronological age or their increased reading proficiency. Regardless of the underlying cause, due to the observed systematic differences, the standard procedure is to control as closely as possible for age and reading proficiency level when designing reading experiments.

There are several psycholinguistic studies that show that also in children, the typicality and plausibility of sentences (Joseph et al., 2008) as well as temporary sentence ambiguity (Traxler, 2002) can be traced in eye movements, suggesting that also other types of comprehension difficulties are reflected in the reading patterns.

Using gaze data to augment models is a recent addition to NLP. Previous approaches that have used gaze data in the context of natural language processing include the work of Barrett et al. (2016), who aim to improve part-of-speech induction with gaze features, Klerke, Goldberg, and Søgaard (2016), where gaze data is used as an auxiliary task in sentence compression, and Klerke et al. (2015), where gaze data is used to evaluate the output of machine translation. The most related work is Klerke, Alonso, and Søgaard (2015) and Gonzalez-Garduño and Søgaard (2017). Klerke, Alonso, and Søgaard (2015) compared gaze from reading original, manually compressed, and automatically compressed sentences. They found that the proportion of regressions to previously read text is sensitive to the differences in human- and computer-induced complexity. Gonzalez-Garduño and Søgaard (2017) show that text readability prediction improves significantly from hard parameter sharing when models try to predict word-based gaze features in a multi-task-learning setup. All of these works, however, use gaze data that was collected under laboratory conditions from skilled, adult readers.

10.3 GAZE DATA

In eye-tracking studies, gaze data is normally sampled under experimental circumstances, where e.g. instructions, location, environment, lighting, participant sampling, textual features, order, duration etc.

Cleaning step	Reading sessions	Unique readers	Read pages	Read words	Misreadings
No cleaning	369	95	3161	73,965	644
Help word activated	366	95	3067	71,911	619
Fixation detection	366	95	3048	64,191	613
Bad calibration	335	87	2865	56,166	565
Marked by teacher	83	44	405	8,681	565

Table 10.1: Dataset size after each cleaning step

are controlled for. Our real-world data, on the contrary, lacks all of these controls. While in controlled, cognitive psychology experiments, fixation durations have proven to systematically correlate with cognitive load (see Rayner et al. (1989) for a review), eye movements from-real world applications have been largely understudied, and specific findings from the literature on controlled data may not apply here or may be swamped by extraneous factors. Further, the often-used statistical tests of significant differences between gaze patterns lose some of their legitimacy when data is retrieved under noisy conditions.

10.3.1 *Data collection and preprocessing*

The data we use in this work is collected in Danish schools using commercial software specifically developed to record and track children's reading development. The system records the eye movements and voice while the children are reading aloud. The teacher can afterwards replay the reading along with the recorded eye movements. The software performs some low-level eye-movement analyses to help the teacher understand how the child processes the text. The teacher can mark which words are erroneously read by the child and later access this and other basic statistics about the reading – see Klerke et al. (2018) for a workflow description. The genre is children's fiction books and the children read contextualized, running text.

As the data is fairly noisy compared to data from laboratory-based eye tracking experiments, we perform thorough cleaning before running any experiments. This cleaning procedure is described below. Table 10.1 contains a summary of the dataset sizes after each cleaning step. Before any cleaning is performed, the dataset contains 369 reading sessions from 95 unique readers. In total it has 3,161 read pages.

HELP WORD ACTIVATED ON PAGE We start by removing all pages where the reader activated the help word function, which dynamically isolates and enlarges a single word on the screen. This dynamic display generates a series of eye movements that do not resemble typical reading activity. This step removes 94 pages.

FIXATION DETECTION We pre-process the raw gaze data by first detecting fixations using a custom implementation of the algorithm of Nyström and Holmqvist (2010). We remove fixations shorter than 40ms and longer than 1.5s.[3] For the calculation of gaze features (see below), we further discard all data points that are not detected as a fixation on text (but instead on images or blank parts of the page). We remove 19 pages where we do not have any fixations on text (e.g. due to the reader just browsing through a book or because of technical issues).

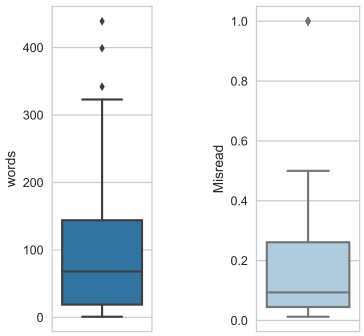
BAD CALIBRATION Prior to reading, the student is prompted to calibrate the eye tracker. In the data used in this study, most reading sessions (91%) attain the best calibration score on a five-point scale, while 6% miss a calibration score. The remaining 3% do not have the best calibration score. We remove everything but the 91% with the best calibration score.

Only parts of the readings have been reviewed and marked for misreadings by a teacher. However, whether a teacher reviewed a reading or not is not explicitly encoded in the data. Thus, if there are no marked misreadings in some session, we do not know whether this is because this reading was not reviewed or because there actually were no errors. We therefore remove all readings without any marked misreadings, as well as any data before the first marked misreading and after the last marked misreading within marked sessions, assuming that everything between these two points has been marked. Twelve cleaned reading sessions only consist of one misread word – everything before and after was removed. See Figure 10.2a for an overview of the distribution of number of words per reading after this cleaning step. This leaves us with the subset of the readings that posed most problems for the subjects. Figure 10.2b shows the distribution of misread words in the cleaned dataset. It is worth noting that since this is not controlled, experimental data, “misread" is not necessarily interpreted equally by all teachers, or even consistently across markings from the same teacher, due to the lack of an annotation protocol. We assume that “misread" means that the pronounced word deviates substantially from the written word. Ultimately, we retain 83 reading sessions from 44 readers with at least one misread word.

10.3.1.1 *Apparatus*

The eye tracker used is a Tobii Eye Tracker 4C with a sample rate of 90 Hz. It is an affordable, consumer eye tracker targeted at gaming. The laptop computers to which the trackers are attached, and which run the software, are provided by the different institutions and

3 Removing short fixations also removes the majority of blinks which presents as a sudden downward-upward pattern of saccades separated by a pause in the signal or a short, falsely detected fixation.



(a) Words per cleaned reading session

(b) Misreading ratio per cleaned session

Figure 10.2: Distributions of total number of words and misreading ratios per session after cleaning.

vary. Screen resolution is locked by the eye tracker software to 1366 x 768, and most systems reportedly run on a 14"–15.6" monitor. The font size is 50pt, which is equivalent to approximately 6mm x-height. Distance between baselines was approximately 18mm with the most commonly used font–otherwise 24mm.

10.3.1.2 *Subjects*

The cleaned dataset contains 44 unique readers with different reading durations. Readers are probably between 5 and 15 years old, which is the official age of students in the Danish schools, but we do not know their exact ages. To control for reading proficiency, we include the texts' readability scores as a feature in all experiments. All students receive extra reading classes, because they struggle with reading. Many of them are probably dyslexic, but we do not have access to this information. Because this is not experimental data, the students will have received different instructions from the teachers. We do not know if they picked the text themselves or for how long they read prior to each recording. They are not necessarily alone in the room, but it is a fair assumption that they all make an effort to read correctly because they are recorded. The data comes from a number of different systems that we were informed is in the range between 10 and 20, but the actual number of schools and teachers is unknown to us. All children and their parents gave consent that the anonymized eye-tracking data may be used for this research.

10.3.2 *Features*

Reading patterns have been shown to be influenced by a number of factors, including textual features and the instructions given to a reader, such as encouraging a specific reading strategy. Readers, or different groups of readers, furthermore display individual reading styles which affect the eye movements (Benfatto et al., 2016). Other factors include the reader's individual skill level, cognitive abilities and mood, among others.

We extract a number of gaze features that have been associated with processing load. Some of our gaze features directly reflect the processing load associated with a word, especially the two correlated measures *total fixation duration* and *number of re-fixations*, but also the *mean fixation duration*. Some gaze features are included to account for preview effects (whether the next or previous word was fixated) as well as the scan path immediately surrounding the word. We split the gaze features into two groups: GAZE (W) for features directly associated with word-level processing and GAZE (C) for features associated with the eye movements on the immediate context of the word. All features are scaled to the $[-1, 1]$ interval.

We further extract a number of basic features that are known to affect gaze features and thus need to be controlled for. These include word length and word frequency (Hyönä and Olson, 1995), but also position in sentence (Rayner, Kambe, and Duffy, 2000) and position on the page have shown to affect reading for adults. We also include a range of linguistic features that we expect to describe word difficulty. All features and feature groups are listed in Table 10.2 and described below.

GAZE FEATURES During reading, the reader performs a series of stable fixations of a couple of hundred milliseconds duration on average. Between fixations, the eyes perform rapid, targeted movements, called *saccades*. All gaze features are computed on the word level and use the application's definition of the area of interest surrounding each word.

For gaze duration, we extract both late and early processing measures. Late measure such as total *fixation duration* and *number of re-fixations* reflect late syntactic and semantic processing in skilled adult reading (Rayner et al., 1989). For children with reading difficulties, we assume these measures to likely reflect processing difficulty.

For the first three passes over a word, we also extract the direction and the word distance of both the ingoing and outgoing saccade.[4]

4 As we removed everything that was not a fixation on text before calculating the gaze features, intermediary non-text fixations may have occurred between text fixations, such as image fixations. We count the last/next fixated *word*. For example, if a word has index 5, and the first pass incoming saccade is from word index 4, we get a feature value of -1 for first pass ingoing.

BASIC	GAZE ON WORD (W)
Is bold	Number of fixations on word
Is italic	First fixation duration
Is lowercase	Mean fixation duration
Is uppercase	Total fixation duration
Has punctuation	Count of passes over the word
Line index on page	Left pupil size
Word index on line	Right pupil size
Page number	Re-fixation counts
Position in sentence (relative)	Fixations in first quarter count
Position in sentence (absolute)	Fixations in second quarter count
Sentence length (characters)	Fixations in third quarter count
Sentence length (words)	Fixations in fourth quarter count
Word index	Relative landing position of first fixation
Sentence index	Relative landing position of last fixation
Word length (characters)	Average character index of fixations
GAZE IN CONTEXT (C)	LINGUISTIC
1*st* pass ingoing saccade dist. and dir.	LIX score for entire text
1*st* pass outgoing saccade dist. and dir.	Previous occurrences of word stem in text
2*nd* pass ingoing saccade dist. and dir.	Previous occurrences of word type in text
2*nd* pass outgoing saccade dist. and dir.	Vowel count
3*rd* pass ingoing saccade dist. and dir.	Character perplexity
3*rd* pass outgoing saccade dist. and dir.	Word frequency
Next word fixated	Universal POS tag
Previous word fixated	

Table 10.2: Overview of the feature groups used in the experiments.

These six features are expected to map the activity around the word and, for example, show whether some word was part of sequential, forward reading or occurred in a series of erratic saccades.

Four features indicate the *landing positions* of fixations in four equally-sized parts of the display width of a word. This captures whether a word, for instance, has three fixations on the last quarter of its display width, which would be atypical and suggest that the reader is struggling with the ending of this word. We further explicitly encode the landing position of the first and last fixation. Note that because of the anatomy of the eye, eye tracking can never be pixel-accurate, but has at least 2° inaccuracy. For short words (or words printed very small, which does not apply for this study) these features may be misleading.

The data also provides pupil sizes for both eyes. It is well known that the pupil dilates as response to external lighting factors, but there is also evidence that the pupil systematically–but on a much

smaller scale–dilates as a response to mental state, emotions or concentration (Beatty, Lucero-Wagoner, et al., 2000). In an experiment collecting pupil size, one would control lighting, which was not possible in the present scenario. For all pupil measures, we subtracted the same side mean of the reading session. We confirmed that all changes larger than 0.6 times the mean were captured when removing short fixations, as they may be caused by the tracker mistaking eyelashes for pupils during blinks.

BASIC FEATURES The basic features span 16 textual and presentational features that are either directly accessible via the system or easily obtainable. They are included in all our experiments and serve as control features for the gaze features because we expect them to explain some of the variance in the gaze features, e.g. reading changes over the course of a line and the course of a sentence (Just and Carpenter, 1980). We further encode the line number a word is located in on a page, as well as its position in that line.

LINGUISTIC FEATURES The linguistic features include the absolute vowel count, which in Danish is highly correlated with the number of syllables. Universal POS tags are obtained from the Danish Polyglot tagger.[5] We also include the provided läsbarhetsindex (LIX) (Björnsson, 1968), a Swedish readability metric (commonly also applied to Danish) that considers the mean sentence length and the ratio of long words (more than 6 characters). The log word probability is estimated from a language model we train on the entire Danish Wikipedia (downloaded in November 2017) using KenLM (Heafield, 2011). Frequency affects processing load and thus fixation duration for adults as well as dyslexic and neurotypical Finnish children (Hyönä and Olson, 1995), but there is conflicting evidence whether text frequencies from adult text explain variance in children's eye movements (Blythe and Joseph, 2011). Character perplexity is estimated using a 5-gram character language model, also using KenLM on the Danish Wikipedia. The previous occurrence of stems and word types is included as reading time for low-frequency words has shown to decrease on later repeats in a text (Rayner, Raney, and Pollatsek, 1995). We use NLTK's snowball stemmer for Danish.

10.4 MODEL

In preliminary experiments, we observed that the relatively small overall amount of data, as well as the low fraction of positive instances, caused significant variation between repeated random restarts of various classification algorithms. We thus approach the task of predicting misreadings from gaze with ensemble methods, training N

5 http://polyglot.readthedocs.io

FEATURE GROUP	F_1	
BASIC	18.78	†
+ GAZE (W)	40.50	*
+ GAZE (C)	18.49	†
+ LINGUISTIC	19.24	†
+ GAZE (W) + GAZE (C)	**41.19**	*
+ GAZE (W) + LINGUISTIC	41.08	*
+ GAZE (W) + LINGUISTIC	18.65	†
All features	40.42	*

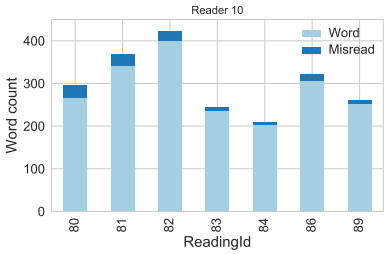
Table 10.3: Performance across feature groups for Experiment 1. Scores are averaged F_1 over ten cross-validation folds. Using an independent t-test, * and † indicate results from ten cross validation rounds significantly different from BASIC and the best feature combination BASIC + GAZE(W) + GAZE(C), respectively.

classifiers independently on the same data and letting them vote on the instances in a held-out development set. Using this development set, we then optimize a threshold t, which is the fraction of the number of classifiers that need to cast a positive vote on an item before we accept it as such.

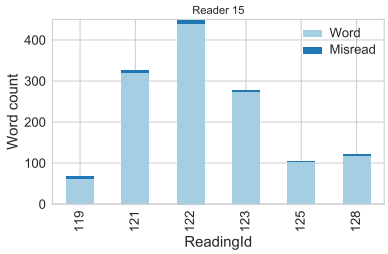
All of our ensembles consist of 10 random forest classifiers and 10 feed-forward neural networks. The random forests, in turn, consist of 100 trees that create splits based on Gini impurity (Breiman, 2001). The neural network models are implemented in Pytorch and trained with the Adam algorithm (Kingma and Ba, 2014), with an initial learning rate of $3 \cdot 10^{-4}$ and a dropout rate of 0.2 on the hidden layers, whose number and sizes we vary in our experiments. We further employ early stopping, monitoring the loss on the development set with a patience of 30 steps.

10.4.1 *Multi-task learning for cross-user knowledge transfer*

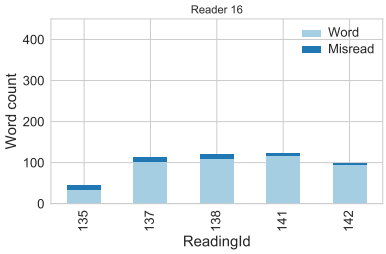
One of the central questions we investigate in this paper is to what degree gaze patterns for misread words vary between readers, and whether we can learn to transfer knowledge about predictors of misreadings between readers. We address these questions in the experiments reported in Section 10.5.2, for which we use a multi-task learning (MTL) model that employs hard parameter sharing. MTL has received significant attention in the natural language processing community over the past years (see Bjerva (2017) for a review). One of the most intriguing properties of MTL is that it allows for the transfer



(a) Reader 10



(b) Reader 15



(c) fig:16wordsmisread

Figure 10.3: Words and misreading counts for readings of three readers in cross-user experiment

of knowledge between different tasks and datasets, which has been investigated and exploited in a growing number of works (Bingel and Søgaard, 2017; Klerke, Goldberg, and Søgaard, 2016; Martínez Alonso and Plank, 2017), including work on the identification of complex words (Bingel and Bjerva, 2018).

In this work, we view the different readers as different *tasks*, motivated by Bingel and Bjerva (2018), who interpret different languages as different tasks for cross-lingual complex word identification. We define a feed-forward neural network model with one output layer per reader, all of which are dense projections from a shared hidden layer. In this framework, each training step consists of flipping a coin to sample any of the tasks and retrieving a batch of training data for this task. This batch is then used to optimize both the shared and the respective task-specific parameters. For a detailed definition of the model, see Bingel and Bjerva (2018).

10.5 EXPERIMENTS

10.5.1 *Experiment 1: Across entire dataset*

As a first experiment, we investigate the performance of our models and the predictiveness of the individual feature groups through 10-fold cross validation across the entire dataset. At each fold, we re-

UserId	Number of reading sessions	Words per reading		Thereof misread	
		Mean	std.dev.	Mean	Std.dev.
10	7	285.9	67.5	16.6	9.9
15	6	219.2	148.1	5.0	2.3
16	5	91.6	32.7	8.0	3.1

Table 10.4: Statistics of (misread) words in sessions for the three readers with most readings.

serve one tenth of the data for testing and another tenth to monitor validation loss of the network as the early stopping criterion.

Note that we split the data randomly and do not stratify the cross-validation splits in any way. In conjunction with the strong class imbalance, this means that we are likely to encounter very different class distributions across splits. This setup may generally lead to lower performance scores, likely with greater variance. However, this was a deliberate choice as we cannot assume a consistent class distribution across train and test set in the real world, or in fact hardly any prior knowledge with regards to class distribution in the test set. Random splitting also means that data from the same *reading* will likely be distributed across train and test partitions for a certain cross-validation iteration.

We perform a first baseline experiment with only the basic features that we list in Table 10.2. On top of this baseline feature set, we perform further experiments, incorporating all combinations over the other feature groups. The results we present in Table 10.3 are based on the best respective model architecture for each feature combination, evaluated via the average over validation splits.[6]

10.5.2 *Experiment 2: Cross-reader prediction*

WITHOUT READER'S OWN DATA In a second experiment, we are interested in how well our model can predict misreadings for specific readers. For this, we identify the three readers with most reading sessions and perform a range of experiments, testing our models on the readings of each of these readers after training them on all other data. We denote the three most active readers by their unique, anonymized IDs as they appear in the dataset: 10, 15 and 16. These readers have 7, 6 and 5 recorded and marked readings, respectively, and we present statistics on these readings in Table 10.4 and Figure 10.3. As in the previous experiment, we optimize our model through cross valida-

6 To address the variation in input dimensionality as we consider different feature group combinations, we train models with different architectures: (i) a single hidden layer with 20 units, (ii) two hidden layers with 20 units each, and (iii) a single hidden layer with 40 units.

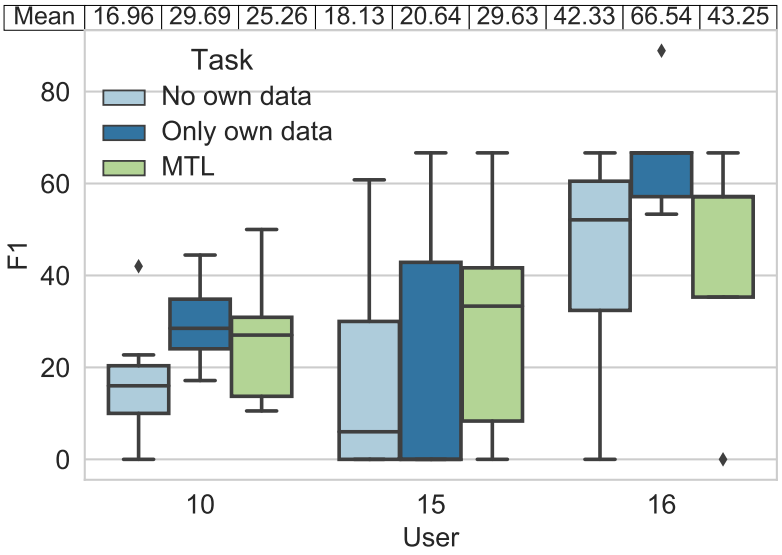


Figure 10.4: F_1 score distributions across test readings for each of the three readers with most sessions for three tasks.

tion to tune hyperparameters and perform early stopping. We report test data results for the model with optimal validation performance in Figure 10.4, broken down into each reader's different sessions.

LEARNING FROM READER'S OWN DATA Complementing the setup above, we now investigate how data from the same reader, but from different reading sessions, can inform our models. Therefore, we further perform cross-validation experiments across each reader's sessions. More concretely, for a reader with n marked readings, we perform n-fold cross validation, holding out one reading a time as a test set and another to monitor validation loss for early stopping of the neural model, while training on the remaining $n - 2$ readings.

MTL As outlined in Section 10.4.1, we now view readers as tasks in an MTL model. For each of the three readers identified above and for each test reading, we train an ensemble whose neural MTL models define two outputs: one for the reader in question and one combined output for all other readers in the entire dataset. The random forest classifiers are trained on all remaining data except the held-out validation and test readings.

10.6 RESULTS AND DISCUSSION

From Experiment 1, we observe that gaze features of the target word itself contribute strongly to model improvements over the baseline of textual features (see Table 10.3). Contextual gaze features and linguistic features do so to a lesser degree. The best feature group combi-

nation consists of the basic features and both gaze feature groups. Adding the linguistic features to this seems to slightly dilute the model.

The results from Experiment 2 in Figure 10.4 show that, at least for these three readers, there is a considerable degree of specificity attested in the reading patterns of misread words: in the scenario where we learn only from other users' gaze patterns (shown in light blue), performance is generally worse than for the other approaches. The high degree of reader specificity is also reflected in the comparison between learning just across a single user's readings and a multi-task setup that also considers other readers. Here, we observe that the former attains higher mean F_1 scores across readings for readers 10 and 16, although MTL is superior to the single-task setup for reader 15. Another observation is that misreadings can generally be predicted much better for reader 16 than for the other readers, which may in part be due to the higher ratio of misread words in these readings.

As especially our cross-reader experiments show, there is reason to believe that the manifestations of misreadings in gaze differ strongly between these readers. However, since we do not have information on the individual readers' age or general reading proficiency, we cannot confidently conclude whether the better stability of within-user experiments attested in Figure 10.4 is due to reader-specific idiosyncrasies or group-internal patterns (which would be supported by evidence that readers 10 and 16 were more atypical readers than others in the present dataset). We find some support for the latter hypothesis in literature describing children's reading development, which identifies a range of patterns common to young and low-proficiency readers. These patterns include longer and more frequent fixations, shorter saccadic amplitude and more regressions – all of which are also associated with comprehension difficulties, see Blythe and Joseph (2011) for a review. The presence of group-internal patterns is further supported by the observation that we are still able to successfully transfer knowledge about readings patterns between users in some cases, increasing performance for the readings of user 15.

One disadvantage of noisy, real-world data is that we do not know to what degree similarities and differences in the data, as well as our results, are influenced by chance, or whether they will generalize to other gaze data. The fact that many parameters are outside of our control and also outside of our knowledge means that we cannot describe certain biases in the data (such as age or reading skill) and consider them as causes for statistical variations in model performance.

10.7 CONCLUSION

This paper presented first work in the automatic prediction of reading errors in children with dyslexia and other reading difficulties using

real-world gaze data. We showed that despite the noisy conditions under which this data was obtained, features we extract from the gaze patterns are predictive of reading mistakes among children. Besides the immediate application in automating some parts of reading teaching, this could be exploited in personalized text simplification, where gaze could be used as feedback to the system.

Our experiments further show that while gaze patterns for misreadings seem to be largely specific to individual readers or groups of readers, we can successfully use MTL to transfer knowledge between readers at least in some cases. Note also that we have very little knowledge of the age and general proficiency of specific readers, including those investigated in our MTL experiments, and we expect that our MTL approach can be much more successful between more similar readers.

ACKNOWLEDGMENTS

The authors wish to thank the children using EyeJustRead, as well as their parents, for giving us permission to use their data. We are also grateful to Emil Juul Jensen and Janus Askø Madsen for supplying us with the data. We further acknowledge valuable comments by Anders Søgaard as well as by the anonymous reviewers. We acknowledge the support by Trygfonden.

Part V

CLOSING REMARKS

11 CONCLUSION AND FUTURE PERSPECTIVES

I have presented eight studies each of which contributes uniquely to a new subfield of NLP that uses data sources containing the cognitive processing signal of text. The studies draw on well-documented conclusions from the field of psycholinguistics, where a wide range of word properties are shown to be reflected in word-level eye-tracking metrics. This knowledge is applied to established NLP tasks.

The key question of this thesis is how data reflecting the human processing of text can benefit NLP.

In all studies evaluating NLP models, the human metrics consistently outperform unmodified baselines and/or feature-enriched baselines. Together these studies provide substantial evidence that the human text processing signal can be used to improve NLP. The supervised models of this thesis evaluate on the following tasks: POS tagging, dependency parsing, sentiment classification, grammatical error detection, detection of abusive language, and prediction of misreadings. The weakly supervised models evaluate on POS induction and chunk induction. This should not be considered the final set of tasks where human data reflecting text processing can help. On the contrary, more work should be done to validate this and uncover the full spectrum.

Below I will provide answers to the questions, posed in Section 1.2.

To what extent can the human processing signal for a broad range of categories be extracted from the eye movements of a reader and be used for POS tagging/syntactic parsing/POS induction?

Five studies in this thesis show that we can successfully extract the word-level processing signal from eye-tracking data and use it for syntax or word class prediction. The pilot experiments in Part ii show that improving supervised POS tagging and supervised parsing with gaze is possible, even across domains. In Chapter 6 and Chapter 7, we successfully use eye-movement features to improve weakly supervised POS induction in a type-constrained SHMM-ME for English and French, respectively. For both languages, type-level averaged features outperform token-level features. The next step will be to obtain eye-tracking data for low-resource languages and confirm this conclusion for these languages.

Our results indicate that models should include a broad set of features reflecting both early and late word processing as well as gaze context features. Chapter 6 tries different gaze and baseline feature

groups and finds that the best gaze feature group contain total fixation duration, mean fixation duration, *n* fixations and fixation probability. But this group alone is beat by frequency and word length baselines. Only when gaze feature groups are combined, is the baseline significantly beaten. Also results from the two pilot experiments in Part ii suggest that the processing signal for word classes and syntax is distributed over many gaze features.

To what extent does the processing signal transfer from one language to another related language for POS induction?

The results in Chapter 7 suggest that the syntactic processing signal from one language to some extent transfers to another. We use the signal from native English speakers to significantly improve POS induction for French but when training on French and testing on English, the minor improvement was not significant. In other words, the way English native speakers process English word classes is similar to the way French native speakers process French word classes to an extent where having eye-tracking data from a related language is better that not having any. This is potentially useful for low-resource languages. Future work should uncover the extent of this phenomenon including the nature of the relatedness between languages for this transfer to happen. Since it was not possible to get a significant improvement for French to English, these two languages may almost be too different, even with comparable eye-tracking datasets. The asymmetry could also be attributed to differences in Wiktionary quality, where the English had a better fit for the data.

How will gaze data support POS induction when combined with other data sources reflecting human text processing, such as features from keystroke logs and acoustic features?

In Chapter 8, we combine the gaze data from the best model from Chapter 6 with four partially overlapping datasets of human text processing as well as pre-trained word embeddings. We find that CCA was the best method for combining sources, and that the best model for POS induction combined Dundee gaze features with pre-trained word embeddings. But also keystroke and acoustic features contributed to models that significantly outperformed the baseline. Combining data sources may lead to more robust models and could also be a pragmatic approach for low-resource languages. Since the data sources in this study were of unequal size, it is difficult to determine which source was more valuable for the tasks. Keystroke and acoustic features were the two smallest datasets, so their potential should be explored further. Technology for collecting these data sources is also mature enough to collect data on a larger scale.

Chapter 8 is the first study in this thesis to only use type-level averaged data. Since we did not need human data for the test set, we are therefore able to evaluate our models on established NLP corpora, instead of eye-tracking corpora. The results in Chapter 6 and Chapter 7 showed that type-level averaged features are better than token-level features. Obtaining type-level features is more resource-efficient than obtaining data reflecting human text processing data at test time. Type-level features should be considered an obvious choice for future work.

To what extent can we use gaze features to guide the attention of a RNN for sequence classification?

Chapter 9 presents another way of including eye-tracking data for NLP, that does not require human data at test time. Gaze durations as continuous values were predicted in a bi-LSTM where the prediction of gaze was an auxiliary task which regularized the main task, thus serving as an inductive bias. We obtained consistent improvements over both a feature-enriched baseline and an unmodified baseline for a wide range of established NLP tasks: grammatical error detection, sentiment classification, and detection of abusive language. Using human data as inductive bias for NLP models seems like a promising direction of research.

How can we model noisy real-word gaze data?

It is an interesting finding that human data collected in lab experiments can contribute to a wide range of NLP tasks but for this idea to really fly, we need to show that human data can contribute equally well on a larger scale. Chapter 10 present the first attempt to model real-world reading data. Although it is a small real-world dataset, it introduces challenges related to noise, bad calibration, and missing information, which makes modelling more difficult than it would have been using laboratory data. We do, however, find that it is possible to achieve good performance when predicting misreadings.

11.1 LIMITATIONS AND CHALLENGES FOR FUTURE WORK

There is general evidence, for example by Osaka (1989), that characteristics of the writing system influence eye movements. All studies in this thesis use reading of alphabet characters. The conclusions drawn at the end should be considered valid only for the studied languages. Although some eye-tracking reading features do seem to generalize across related languages and it seems intuitively likely that children may exhibit more or less similar misreading behaviour in at least related languages, it is beyond the scope of this thesis to explore to what

extent this holds. However, the fact that we can use the same methodology for several languages, and even show some transfer from one language to another, means that it would make perfect sense for future work to approach other languages and even writing systems using the same methodology that we propose.

Real-world human data sources present privacy issues beyond those of laboratory data. Real-world data will contain information about what people actually choose to read, say and type, and such data is highly sensitive on an individual level and maybe also on a group-level. It should be gathered and stored ethically. Unless this challenge is solved adequately, it could prevent human data from being used without legal consequences or result in data being abused which in turn would discourage people from providing data.

Larger quantities of real-world data will also reduce the risk of overfitting models to the specific data from one experiment, but introduce other issues such as noise, missing data, domain biases, and privacy issues. In the presented studies, only data from skilled, adult reading is used, except in Chapter 10 that use data from children with reading difficulties. But real-world data will contain data from neurotypical people mixed with data from non-neurotypical people, just to mention one factor that should not be logged according to ethical issues, but will bias data heavily. This will be a challenge for future work.

Part VI

APPENDIX

A

GAZE FEATURES

- First fixation duration on every word
- Fixation probability
- Mean fixation duration per sentence
- Mean fixation duration per word
- Next fixation duration
- Next word fixation probability
- Probability to get 1^{st} fixation
- Probability to get 2^{nd} fixation
- Previous fixation duration, previous word fixation probability
- Re-read probability
- Reading time per sentence normalized by word count
- Share of fixated words per sentence
- Time percentage spent on this word out of total sentence reading time
- Total fixation duration per word
- Total regression from word duration
- Total duration of regressions to word
- *n* fixations on word
- *n* fixations per sent normalized by token count
- *n* long regressions from word
- *n* long regressions per sentence normalized by token count
- *n* long regressions to word
- *n* re-fixations on word
- *n* re-fixations per sentence normalized by token count
- *n* regressions from word
- *n* regressions per sentence normalized by token count
- *n* regressions to word

BIBLIOGRAPHY

Abdelali, Ahmed, Nadir Durrani, and Francisco Guzmán (2016). "iAppraise: A Manual Machine Translation Evaluation Environment Supporting Eye-tracking." In: *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pp. 17–21.

Abeillé, Anne, Lionel Clément, and François Toussenel (2003). "Building a treebank for French." In: *Treebanks*, pp. 165–187.

Agić, Željko et al. (2015). *Universal Dependencies 1.1*. LINDAT/CLARIN digital library, Institute of Formal and Applied Linguistics, Charles University in Prague. URL: http://hdl.handle.net/11234/LRT-1478.

Alkhouli, Tamer, Gabriel Bretschner, Jan-Thorsten Peter, Mohammed Hethnawi, Andreas Guta, and Hermann Ney (2016). "Alignment-Based Neural Machine Translation." In: *Proceedings of the First Conference on Machine Translation*, pp. 54–65.

Augereau, Olivier, Kai Kunze, Hiroki Fujiyoshi, and Koichi Kise (2016). "Estimation of English skill with a mobile eye tracker." In: *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, pp. 1777–1781.

Baaijen, Veerle M, David Galbraith, and Kees de Glopper (2012). "Keystroke analysis: Reflections on procedures and measures." In: *Written Communication* 29.3, pp. 246–277.

Baayen, Harald R, Ton Dijkstra, and Robert Schreuder (1997). "Singulars and plurals in Dutch: Evidence for a parallel dual-route model." In: *Journal of Memory and Language* 37.1, pp. 94–117.

Barrett, Maria, Željko Agić, and Anders Søgaard (2015). "The Dundee Treebank." In: *The 14th International Workshop on Treebanks and Linguistic Theories (TLT 14)*, pp. 242–248.

Barrett, Maria, Frank Keller, and Anders Søgaard (2016). "Cross-lingual transfer of correlations between parts of speech and gaze features." In: *Proceedings of the 26th International Conference on Computational Linguistics (COLING)*, pp. 1330–1339.

Barrett, Maria and Anders Søgaard (2015a). "Reading behavior predicts syntactic categories." In: *Proceedings of the nineteenth conference on computational natural language learning (CoNLL)*, pp. 345–249.

Barrett, Maria and Anders Søgaard (2015b). "Using reading behavior to predict grammatical functions." In: *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning (CogACL)*, pp. 1–5.

Barrett, Maria, Joachim Bingel, Frank Keller, and Anders Søgaard (2016). "Weakly Supervised Part-of-speech Tagging Using Eye-tracking Data." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vol. 2, pp. 579–584.

Bates, Douglas, Martin Mächler, Ben Bolker, and Steve Walker (2015). "Fitting Linear Mixed-Effects Models Using lme4." In: *Journal of Statistical Software* 67.1, pp. 1–48. DOI: 10.18637/jss.v067.i01.

Bauman, Peter (2013). "Syntactic category disambiguation within an architecture of human language processing." In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 35. 35, pp. 1833–1838.

Beatty, Jackson, Brennis Lucero-Wagoner, et al. (2000). "The pupillary system." In: *Handbook of psychophysiology* 2, pp. 142–162.

Benfatto, Mattias Nilsson, Gustaf Öqvist Seimyr, Jan Ygge, Tony Pansell, Agneta Rydberg, and Christer Jacobson (2016). "Screening for dyslexia using eye tracking during reading." In: *PloS one* 11.12, e0165508.

Berg-Kirkpatrick, Taylor, Alexandre Bouchard-Cote, John DeNero, and Dan Klein (2010). "Painless Unsupervised Learning with Features." In: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 582–590.

Berzak, Yevgeni, Boris Katz, and Roger Levy (2018). "Assessing Language Proficiency from Eye Movements in Reading." In: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 1986–1996.

Berzak, Yevgeni, Chie Nakamura, Suzanne Flynn, and Boris Katz (2017). "Predicting Native Language from Gaze." In: *ACL*.

Bingel Joachim, Gustavo H. Paetzold and Anders Søgaard (2018). "Lexi: a tool for adaptive, personalized text simplification." In: *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*.

Bingel, Joachim, Maria Barrett, and Anders Søgaard (2016). "Extracting token-level signals of syntactic processing from fMRI-with an application to POS induction." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vol. 1, pp. 747–755.

Bingel, Joachim and Johannes Bjerva (2018). "Cross-lingual complex word identification with multitask learning." In: *Proceedings of the Complex Word Identification Shared Task at the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. New Orleans, United States: Association for Computational Linguistics.

Bingel, Joachim and Anders Søgaard (2017). "Identifying beneficial task relations for multi-task learning in deep neural networks."

In: *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*. Vol. 2, pp. 164–169.

Bjerva, Johannes (2017). "One Model to Rule them all: Multitask and Multilingual Modelling for Lexical Analysis." PhD thesis. University of Groningen.

Björnsson, Carl Hugo (1968). *Läsbarhet*. Liber.

Blythe, Hazel I and Holly SSL Joseph (2011). "Children's eye movements during reading." In: *The Oxford Handbook of Eye Movements*. Ed. by I. D. Gilchrist S. P. Liversedge and S. Everling, pp. 643–662.

Bohnet, Bernd (2010). "Very high accuracy and fast dependency parsing is not a contradiction." In: *Proceedings of the 23rd international conference on computational linguistics (COLING*, pp. 89–97.

Breiman, Leo (2001). "Random forests." In: *Machine learning* 45.1, pp. 5–32.

Brent, M. R. and J. M Siskind (2001). "The role of exposure to isolated words in early vocabulary development." In: *Cognition* 81, pp. 31–44.

Britz, Denny, Melody Y. Guan, and Minh-Thang Luong (2017). "Efficient Attention using a Fixed-Size Memory Representation." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 392–400.

Calmettes, Guillaume, Gordon B Drummond, and Sarah L Vowler (2012). "Making do with what we have: use your bootstraps." In: *The Journal of physiology* 590.15, pp. 3403–3406.

Carpenter, Patricia A and Marcel Adam Just (1983). "What your eyes do while your mind is reading." In: *Eye movements in reading: Perceptual and language processes*, pp. 275–307.

Christodoulopoulos, Christos, Sharon Goldwater, and Mark Steedman (2010). "Two Decades of Unsupervised POS Induction: How Far Have We Come?" In: *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pp. 575–584.

Clifton, Charles, Adrian Staub, and Keith Rayner (2007). "Eye movements in reading words and sentences." In: *Eye Movements: A Window on Mind and Brain*. Amsterdam, The Netherlands: Elsevier, pp. 341–371.

Coleman, Meri and Ta Lin Liau (1975). "A computer readability formula designed for machine scoring." In: *Journal of Applied Psychology* 60.2, pp. 283–284.

Collins, Michael (2002). "Discriminative training methods for Hidden Markov Models." In: *Proceedings of the ACL-02 conference on Empirical methods in natural language processing (EMNLP)*, pp. 1–8.

Consortium, British National Corpus et al. (2007). *British National Corpus version 3*.

Cop, Uschi, Nicolas Dirix, Denis Drieghe, and Wouter Duyck (2017). "Presenting GECO: An eyetracking corpus of monolingual and bilingual sentence reading." In: *Behavior research methods* 49.2, pp. 602–615.

Cordella, L. P., P. Foggia, C. Sansone, and M. Vento (2001). "An Improved Algorithm for Matching Large Graphs." In: *Proceedings of the 3rd IAPR TC-15 Workshop on Graphbased Representations in Pattern Recognition* 17, pp. 1–35.

Culotta, Aron, Andrew McCallum, and Jonathan Betz (2006). "Integrating probabilistic extraction models and data mining to discover relations and patterns in text." In: *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, pp. 296–303.

Das, Abhishek, Harsh Agrawal, Lawrence Zitnick, Devi Parikh, and Dhruv Batra (2017). "Human attention in visual question answering: Do humans and deep networks look at the same regions?" In: *Computer Vision and Image Understanding* 163, pp. 90–100.

Demberg, Vera and Frank Keller (2008). "Data from eye-tracking corpora as evidence for theories of syntactic processing complexity." In: *Cognition* 109, pp. 193–210.

Demuth, Katherine, Jennifer Culbertson, and Jennifer Alter (2006). "Word-minimality, Epenthesis and Coda Licensing in the Early Acquisition of English." In: *Language and Speech* 49.2, pp. 137–173.

Dhillon, Paramveer, Dean Foster, and Lyle Ungar (2015). "Eigenwords: Spectral Word Embeddings." In: *Journal of Machine Learning Research* 16, pp. 3035–3078.

Dong, Daxiang, Hua Wu, Wei He, Dianhai Yu, and Haifeng Wang (2015). "Multi-task learning for multiple language translation." In: *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*. Vol. 1, pp. 1723–1732.

Eyben, Florian et al. (2016). "The Geneva Minimalistic Acoustic Parameter Set (GeMAPS) for Voice Research and Affective Computing." In: *Language and Speech* 7, pp. 190–202.

Faruqui, Manaal and Chris Dyer (2014a). "Community Evaluation and Exchange of Word Vectors at wordvectors.org." In: *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics (ACL) : System Demonstrations*, pp. 19–24.

Faruqui, Manaal and Chris Dyer (2014b). "Improving vector space word representations using multilingual correlation." In: *Proceedings of the 14th Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, pp. 462–471.

Felbo, Bjarke, Alan Mislove, Anders Søgaard, Iyan Rahwan, and Sune. Lehmann (2017). "Using millions of emoji occurrences to pretrain

any-domain models for detecting emotion, sentiment, and sarcasm." In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1615–1625.

Firat, Orhan, Kyunghyun Cho, and Yoshua Bengio (2016). "Multi-Way, Multilingual Neural Machine Translation with a Shared Attention Mechanism." In: *Proceedings of 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 866–875.

Fossum, Victoria and Roger Levy (2012). "Sequential vs. hierarchical syntactic models of human incremental sentence processing." In: *Proceedings of the 3rd workshop on cognitive modeling and computational linguistics*. Association for Computational Linguistics, pp. 61–69.

Foster, Jennifer, Ozlem Cetinoglu, Joachim Wagner, Josef Le Roux, Joakim Nivre, Deirde Hogan, and Josef van Genabith (2011). "From news to comments: Resources and benchmarks for parsing the language of Web 2.0." In: *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)*, pp. 893–901.

Frank, Stefan L (2009). "Surprisal-based comparison between a symbolic and a connectionist model of sentence processing." In: *CogSci*, pp. 1139–1144.

Frank, Stefan L (2010). "Uncertainty reduction as a measure of cognitive processing effort." In: *Proceedings of the 2010 workshop on cognitive modeling and computational linguistics*. Association for Computational Linguistics, pp. 81–89.

Frank, Stefan L and Rens Bod (2011). "Insensitivity of the human sentence-processing system to hierarchical structure." In: *Psychological Science* 22.6, pp. 829–834.

Frank, Stefan L, Irene Fernandez Monsalve, Robin L Thompson, and Gabriella Vigliocco (2013a). "Reading time data for evaluating broad-coverage models of English sentence processing." In: *Behavior Research Methods* 45.4, pp. 1182–1190.

Frank, Stefan L, Leun J Otten, Giulia Galli, and Gabriella Vigliocco (2013b). "Word surprisal predicts N400 amplitude during reading." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*. Vol. 2, pp. 878–883.

Frank, Stefan and Robin Thompson (2012). "Early effects of word surprisal on pupil size during reading." In: *Proceedings of the Annual Meeting of the Cognitive Science Society*. Vol. 34. 34, pp. 1554–1559.

Frazier, Lyn and Keith Rayner (1982). "Making and correcting errors during sentence comprehension: Eye movements in the analysis of structurally ambiguous sentences." In: *Cognitive psychology* 14.2, pp. 178–210.

Frermann, Lea and Michael C. Frank (2017). "Prosodic Features from Large Corpora of Child-Directed Speech as Predictors of the Age of Acquisition of Words." In: *Computing Research Repository (CoRR)*.

Furtner, Marco R, John F Rauthmann, and Pierre Sachse (2009). "Nomen est omen: Investigating the dominance of nouns in word comprehension with eye movement analyses." In: *Advances in Cognitive Psychology* 5, pp. 91–104.

Furtner, Marco R, John F Rauthmann, and Pierre Sachse (2011). "Investigating word class effects in first and second languages." In: *Perceptual and motor skills* 113.1, pp. 87–97.

Ganchev, Kuzman, Keith Hall, Ryan McDonald, and Slav Petrov (2012). "Using search-logs to improve query tagging." In: *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 238–242.

Garrette, Dan and Jason Baldridge (2013). "Learning a part-of-speech tagger from two hours of annotation." In: *Proceedings of the 11th Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pp. 138–147.

Gibson, E and K Ko (1998). "An integration-based theory of computational resources in sentence comprehension." In: *Fourth Architectures and Mechanisms in Language Processing Conference, University of Freiburg, Germany*.

Gibson, Edward (1998). "Linguistic complexity: Locality of syntactic dependencies." In: *Cognition* 68.1, pp. 1–76.

Gibson, Edward (2000). "The dependency locality theory: A distance-based theory of linguistic complexity." In: *Image, language, brain*, pp. 95–126.

Gitari, Njagi Dennis, Zhang Zuping, Hanyurwimfura Damien, and Jun Long (2015). "A lexicon-based approach for hate speech detection." In: *Journal of Multimedia and Ubiquitous Engineering* 10.4, pp. 215–230.

Globerson, Amir and Sam Roweis (2006). "Nightmare at test time: robust learning by feature deletion." In: *Proceedings of the 23rd International Conference on Machine Learning (ICML)*, pp. 353–360.

Gonzalez-Garduño, Ana Valeria and Anders Søgaard (2017). "Using gaze to predict text readability." In: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pp. 438–443.

Gonzalez-Garduno, Ana and Anders Søgaard (2018). "Learning to predict readability using eye-movement data from natives and learners." In: *Proceedings of the Thirty-Second Association for the Advancement of Artificial Intelligence Conference (AAAI)*.

Goodkind, Adam and Andrew Rosenberg (2015). "Muddying The Multiword Expression Waters: How Cognitive Demand Affects Multiword Expression Production." In: *Proceedings of the 11th Workshop on Multiword Expressions*, pp. 87–95.

Green, Matthew J (2014). "An eye-tracking evaluation of some parser complexity metrics." In: *Proceedings of the 3rd Workshop on Pre-*

dicting and Improving Text Readability for Target Reader Populations (PITR), pp. 38–46.

Hahn, Michael and Frank Keller (2016). "Modeling Human Reading with Neural Attention." In: *Proceedings of Empirical Methods in Natural Language Processing (EMNLP)*, pp. 85–95.

Hale, John (2001). "A probabilistic Earley parser as a psycholinguistic model." In: *Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies (NAACL)*. Association for Computational Linguistics, pp. 1–8.

Hara, Tadayoshi, Daichi Mochihashi2 Yoshinobu Kano, and Akiko Aizawa (2012). "Predicting word fixations in text with a CRF model for capturing general reading strategies among readers." In: *Proceedings of the 1st Workshop on Eye-tracking and Natural Language Processing*, pp. 55–70.

Hardoon, David R, John Shawe-Taylor, Antti Ajanki, Kai PuolamÃ, Samuel Kaski, et al. (2007). "Information retrieval by inferring implicit queries from eye movements." In: *Artificial Intelligence and Statistics*, pp. 179–186.

Heafield, Kenneth (2011). "KenLM: Faster and smaller language model queries." In: *Proceedings of the Sixth Workshop on Statistical Machine Translation*. Association for Computational Linguistics, pp. 187–197.

Hochreiter, Sepp and Jürgen Schmidhuber (1997). "Long short-term memory." In: *Neural computation* 9.8, pp. 1735–1780.

Hollenstein, Nora, Nicolas Langer, Andreas Pedroni, Marius Troendle, Jonathan Rotsztejn, and Ce Zhang (2018). *Zurich Cognitive Language Processing Corpus: A simultaneous EEG and eye-tracking resource to analyze the human reading process*. URL: osf.io/q3zws.

Hyönä, Jukka, Raymond Bertram, and Alexander Pollatsek (2004). "Are long compound words identified serially via their constituents? Evidence from an eyemovement-contingent display change study." In: *Memory & Cognition* 32.4, pp. 523–532.

Hyönä, Jukka and Richard K Olson (1995). "Eye fixation patterns among dyslexic and normal readers: Effects of word length and word frequency." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 21.6, pp. 1430–40.

Ibraheem, Samee, Nicholas Altieri, and John DeNero (2017). "Learning an Interactive Attention Policy for Neural Machine Translation." In: *MTSummit*.

Immonen, Sini and Jukka Mäkisalo (2010). "Pauses Reflecting the Processing of Syntactic Units in Monolingual Text Production and Translation." In: *HERMES–Journal of Language and Communication in Business* 23.44, pp. 45–61.

Interagency Committee on Learning Disabilities (1987). *Learning Disabilities: A Report to the U.S. Congress.* Tech. rep. Government Printing Office, Washington DC, U.S.

Ivanko, Stacey L and Penny M Pexman (2003). "Context incongruity and irony processing." In: *Discourse Processes* 35.3, pp. 241–279.

Jaffe, Evan, Cory Shain, and William Schuler (2018). "Coreference and Focus in Reading Times." In: *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL)*, pp. 1–9.

Joseph, Holly SSL, Simon P Liversedge, Hazel I Blythe, Sarah J White, Susan E Gathercole, and Keith Rayner (2008). "Children's and adults' processing of anomaly and implausibility during reading: Evidence from eye movements." In: *Quarterly Journal of Experimental Psychology* 61.5, pp. 708–723.

Juhasz, B. J. and A. Pollatsek (2011). "Lexical influences on eye movements in reading." In: *The Oxford Handbook of Eye Movements*. Ed. by I. D. Gilchrist S. P. Liversedge and S. Everling, 873–893).

Juhasz, Barbara J and Keith Rayner (2003). "Investigating the effects of a set of intercorrelated variables on eye fixation durations in reading." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 29.6, p. 1312.

Just, Marcel A and Patricia A Carpenter (1980). "A theory of reading: From eye fixations to comprehension." In: *Psychological review* 87.4, pp. 109–30.

Kennedy, Alan, Robin Hill, and Joël Pynte (2003). "The Dundee corpus." In: *Proceedings of the European Conference on Eye Movements (ECEM)*.

Kennedy, Alan and Joël Pynte (2005). "Parafoveal-on-foveal effects in normal reading." In: *Vision research* 45.2, pp. 153–168.

Kennedy, Alan, Joël Pynte, Wayne S Murray, and Shirley-Anne Paul (2013). "Frequency and predictability effects in the Dundee Corpus: An eye movement analysis." In: *The Quarterly Journal of Experimental Psychology* 66.3, pp. 601–618.

Kilgarriff, Adam (1995). "BNC database and word frequency lists." In: *Retrieved Dec. 2017*.

Killourhy, Kevin S and Roy A Maxion (2012). "Free vs. transcribed text for keystroke-dynamics evaluations." In: *Proceedings of the 2012 Workshop on Learning from Authoritative Security Experiment Results*. ACM, pp. 1–8.

Kingma, Diederik P and Jimmy Ba (2014). "Adam: A method for stochastic optimization." In: *arXiv preprint arXiv:1412.6980*.

Klerke, Sigrid, Héctor Martínez Alonso, and Anders Søgaard (2015). "Looking hard: Eye tracking for detecting grammaticality of automatically compressed sentences." In: *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pp. 97–105.

Klerke, Sigrid, Yoav Goldberg, and Anders Søgaard (2016). "Improving sentence compression by learning to predict gaze." In: *Proceedings of 14th Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pp. 1528–1533.

Klerke, Sigrid, Sheila Castilho, Maria Barrett, and Anders Søgaard (2015). "Reading metrics for estimating task efficiency with MT output." In: *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning (CogACLL)*, pp. 6–13.

Klerke, Sigrid, Janus Askø Madsen, Emil Juul Jacobsen, and John Paulin Hansen (2018). "Substantiating Reading Teachers with Scanpaths." In:

Kliegl, Reinhold and Ralf Engbert (2005). "Fixation durations before word skipping in reading." In: *Psychonomic Bulletin & Review* 12.1, pp. 132–138.

Kliegl, Reinhold, Ellen Grabner, Martin Rolfs, and Ralf Engbert (2004). "Length, frequency, and predictability effects of words on eye movements in reading." In: *European Journal of Cognitive Psychology* 16.1-2, pp. 262–284.

Konieczny, Lars (2000). "Locality and parsing complexity." In: *Journal of psycholinguistic research* 29.6, pp. 627–645.

Krafka, Kyle, Aditya Khosla, Petr Kellnhofer, Harini Kannan, Suchendra Bhandarkar, Wojciech Matusik, and Antonio Torralba (2016). "Eye tracking for everyone." In: *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 2176–2184.

Kuperman, Victor, Michael Dambacher, Antje Nuthmann, and Reinhold Kliegl (2010). "The effect of word position on eye-movements in sentence and paragraph reading." In: *The Quarterly Journal of Experimental Psychology* 63.9, pp. 1838–1857.

Levy, Roger (2008). "Expectation-based syntactic comprehension." In: *Cognition* 106.3, pp. 1126–1177.

Li, Shen, João Graça, and Ben Taskar (2012). "Wiki-ly supervised part-of-speech tagging." In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 1389–1398.

Lindsey, Jack (2017). "Pre-training Attention Mechanisms." In: *NIPS Workshop on Cognitive Informed Artificial Intelligence*.

Liu, Pengfei, Xipeng Qiu, and Xuanjing Huang (2016). "Deep Multi-Task Learning with Shared Memory." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 118–127.

Loboda, Tomasz D, Peter Brusilovsky, and Jöerg Brunstein (2011). "Inferring word relevance from eye-movements of readers." In: *Proceedings of the 16th international conference on intelligent user interfaces*. ACM, pp. 175–184.

Luke, Steven G and Kiel Christianson (2018). "The Provo Corpus: A large eye-tracking corpus with predictability norms." In: *Behavior research methods* 50.2, pp. 826–833.

Luong, Minh-Thang, Quoc V. Le, Ilya Sutskever, Oriol Vinyals, and Lukasz Kaiser (2016). "Multi-task sequence-to-sequence learning." In: *International Conference on Learning Representations (ICLR)*, pp. 1–10.

Luzzatti, Claudio, Rossella Raggi, Giusy Zonca, Caterina Pistarini, Antonella Contardi, and Gian-Domenico Pinna (2002). "Verb–noun double dissociation in aphasic lexical impairments: The role of word frequency and imageability." In: *Brain and language* 81.1-3, pp. 432–444.

van der Maaten, Laurens and Geoffrey Hinton (2008). "Visualizing data using t-SNE." In: *Journal of Machine Learning Research* 9.Nov, pp. 2579–2605.

MacWhinney, Brian (2000). *The CHILDES project: Tools for analyzing talk*. Third Edition. Hillsdale, NJ, USA: Lawrence Erlbaum Associates.

von der Malsburg, Titus and Shravan Vasishth (2011). "What is the scanpath signature of syntactic reanalysis?" In: *Journal of Memory and Language* 65.2, pp. 109–127.

von der Malsburg, Titus, Reinhold Kliegl, and Shravan Vasishth (2015). "Determinants of scanpath regularity in reading." In: *Cognitive science* 39.7, pp. 1675–1703.

Manning, Christopher D (2011). "Part-of-speech tagging from 97% to 100%: is it time for some linguistics?" In: *International Conference on Intelligent Text Processing and Computational Linguistics*. Springer, pp. 171–189.

Marcus, Mitchell, Mary Marcinkiewicz, and Beatrice Santorini (1993). "Building a large annotated corpus of English: the Penn Treebank." In: *Computational Linguistics* 19.2, pp. 313–330.

de Marneffe, Marie-Catherine, Timothy Dozat, Natalia Silveira, Katri Haverinen, Filip Ginter, Joakim Nivre, and Christopher D. Manning (2014). "Universal Stanford dependencies: A cross-linguistic typology." In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 4585–4592.

Martínez Alonso, Héctor and Barbara Plank (2017). "When is multitask learning effective? Semantic sequence prediction under varying data conditions." In: *15th Conference of the European Chapter of the Association for Computational Linguistics*.

Matsuhashi, Ann (1981). "Pausing and planning: The tempo of written discourse production." In: *Research in the Teaching of English*, pp. 113–134.

Matthies, Franz and Anders Søgaard (2013). "With Blinkers on: Robust Prediction of Eye Movements across Readers." In: *Proceed-*

ings of the 2013 Conference on Empirical Methods in Natural Language Processing (EMNLP). Seattle, Washington, USA, pp. 803–806.

McConkie, George W and Keith Rayner (1976). "Asymmetry of the perceptual span in reading." In: *Bulletin of the psychonomic society* 8.5, pp. 365–368.

McDonald, Scott A and Richard C Shillcock (2003). "Low-level predictive inference in reading: The influence of transitional probabilities on eye movements." In: *Vision Research* 43.16, pp. 1735–1751.

Meinshausen, Nicolai and Peter Bühlmann (2010). "Stability selection." In: *Journal of the Royal Statistical Society: Series B (Statistical Methodology)* 72.4, pp. 417–473.

Mishra, Abhijit, Pushpak Bhattacharyya, and Michael Carl (2013). "Automatically predicting sentence translation difficulty." In: *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (ACL)*. Vol. 2, pp. 346–351.

Mishra, Abhijit, Kuntal Dey, and Pushpak Bhattacharyya (2017). "Learning Cognitive Features from Gaze Data for Sentiment and Sarcasm Classification using Convolutional Neural Network." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vol. 1, pp. 377–387.

Mishra, Abhijit, Diptesh Kanojia, and Pushpak Bhattacharyya (2016). "Predicting Readers' Sarcasm Understandability by Modeling Gaze Behavior." In: *Proceedings of the 30th Association for the Advancement of Artificial Intelligence Conference (AAAI)*, pp. 3747–3753.

Mishra, Abhijit, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya (2016a). "Harnessing cognitive features for sarcasm detection." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pp. 1095–1104.

Mishra, Abhijit, Diptesh Kanojia, Seema Nagar, Kuntal Dey, and Pushpak Bhattacharyya (2016b). "Leveraging cognitive features for sentiment analysis." In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pp. 156–166.

Mishra, Abhijit, Srikanth Tamilselvam, Riddhiman Dasgupta, Seema Nagar, and Kuntal Dey (2018). "Cognition-Cognizant Sentiment Analysis with Multitask Subjectivity Summarization based on Annotators' Gaze Behavior." In: *Proceedings of the Thirty-Second Association for the Advancement of Artificial Intelligence Conference on Artificial Intelligence*.

Mitchell, Jeff, Mirella Lapata, Vera Demberg, and Frank Keller (2010). "Syntactic and semantic factors in processing difficulty: An integrated measure." In: *Proceedings of the 48th annual meeting of the Association for Computational Linguistics (ACL*, pp. 196–206.

Monsalve, Irene Fernandez, Stefan L Frank, and Gabriella Vigliocco (2012). "Lexical surprisal as a general predictor of reading time." In: *Proceedings of the 13th Conference of the European Chapter of*

the Association for Computational Linguistics (EACL. Association for Computational Linguistics, pp. 398–408.

New, Boris, Marc Brysbaert, Juan Segui, Ludovic Ferrand, and Kathleen Rastle (2004). "The processing of singular and plural nouns in French and English." In: *Journal of Memory and Language* 51.4, pp. 568–585.

Nilsson, Matthias and Joakim Nivre (2009). "Learning where to look: Modeling eye movements in reading." In: *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL)*, pp. 93–101.

Nivre, Joakim, Johan Hall, Jens Nilsson, Atanas Chanev, Gülsen Eryigit, Sandra Kübler, Svetoslav Marinov, and Erwin Marsi (2007). "MaltParser." In: *Natural Language Engineering* 13.2, pp. 95–135.

Nyström, Marcus and Kenneth Holmqvist (2010). "An adaptive algorithm for fixation, saccade, and glissade detection in eyetracking data." In: *Behavior research methods* 42.1, pp. 188–204.

Osaka, Naoyuki (1989). "Eye fixation and saccade during kana and kanji text reading: Comparison of English and Japanese text processing." In: *Bulletin of the Psychonomic Society* 27.6, pp. 548–550.

Owoputi, Olutobi, Brendan O'Connor, Chris Dyer, Kevin Gimpel, Nathan Schneider, and Noah A Smith (2013). "Proceedings of the 2013 conference of the North American chapter of the Association for Computational Linguistics (ACL): human language technologies." In: *NAACL*, pp. 380–391.

Paetzold, Gustavo and Lucia Specia (2016). "Semeval 2016 task 11: Complex word identification." In: *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*, pp. 560–569.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu (2002). "BLEU: a method for automatic evaluation of machine translation." In: *Proceedings of the 40th annual meeting on association for computational linguistics*. Association for Computational Linguistics, pp. 311–318.

Pate, John K and Sharon Goldwater (2011). "Unsupervised syntactic chunking with acoustic cues: computational models for prosodic bootstrapping." In: *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*. Association for Computational Linguistics, pp. 20–29.

Pate, John K and Sharon Goldwater (2013). "Unsupervised dependency parsing with acoustic cues." In: *Transactions of the Association for Computational Linguistics (TACL)* 1, pp. 63–74.

Pennington, Jeffrey, Richard Socher, and Christopher Manning (2014). "Glove: Global vectors for word representation." In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pp. 1532–1543.

Petrov, Slav, Dipanjan Das, and Ryan McDonald (2011). "A universal part-of-speech tagset." In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 2089–2094.

Plank, Barbara (2016a). "Keystroke dynamics as signal for shallow syntactic parsing." In: *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*, pp. 609–618.

Plank, Barbara (2016b). "What to do about non-standard (or non-canonical) language in NLP." In: *Proceedings of the Conference on Natural Language Processing (KONVENS)*, pp. 13–20.

Plank, Barbara, Yoav Goldberg, and Anders Søgaard (2016). "Multilingual part-of-speech tagging with bidirectional long short-term memory models and auxiliary loss." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 412–418.

Pollatsek, Alexander, Shmuel Bolozky, Arnold D Well, and Keith Rayner (1981). "Asymmetries in the perceptual span for Israeli readers." In: *Brain and language* 14.1, pp. 174–180.

Pynte, Joel and Alan Kennedy (2006). "An influence over eye movements in reading exerted from beyond the level of the word: Evidence from reading English and French." In: *Vision Research* 46.22, pp. 3786–3801.

Pynte, Joël and Alan Kennedy (2007). "The influence of punctuation and word class on distributed processing in normal reading." In: *Vision Research* 47.9, pp. 1215–1227.

Pynte, Joel, Boris New, and Alan Kennedy (2009). "On-line contextual influences during reading normal text: The role of nouns, verbs and adjectives." In: *Vision research* 49.5, pp. 544–552.

R Core Team (2015). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing. Vienna, Austria. URL: http://www.R-project.org/.

Raudonis, Vidas, Gintaras Dervinis, Andrius Vilkauskas, Agne Paulauskaite Taraseviciene, and Gintare Kersulyte-Raudone (2013). "Evaluation of human emotion from eye motions." In: *Evaluation* 4.8.

Rauzy, Stéphane and Philippe Blache (2012). "Robustness and processing difficulty models. a pilot study for eye-tracking data on the French Treebank." In: *Proceedings of the 1st Eye-Tracking and NLP workshop*.

Rayner, Keith (1977). "Visual attention in reading: Eye movements reflect cognitive processes." In: *Memory & Cognition* 5.4, pp. 443–448.

Rayner, Keith (1998). "Eye movements in reading and information processing: 20 years of research." In: *Psychological bulletin* 124.3, pp. 372–422.

Rayner, Keith and Susan A Duffy (1986). "Lexical complexity and fixation times in reading: Effects of word frequency, verb complexity, and lexical ambiguity." In: *Memory & cognition* 14.3, pp. 191–201.

Rayner, Keith and Susan A. Duffy (1988). "On-line comprehension processes and eye movements in reading." In: *Reading research: Advances in theory and practice*. Ed. by G. E. MacKinnon M. Daneman and T. G. Waller (Eds.) New York, NY, USA: Academic Press, pp. 13–66.

Rayner, Keith and Lyn Frazier (1989). "Selection mechanisms in reading lexically ambiguous words." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 15.5, p. 779.

Rayner, Keith, Gretchen Kambe, and Susan A Duffy (2000). "The effect of clause wrap-up on eye movements during reading." In: *The Quarterly Journal of Experimental Psychology Section A* 53.4, pp. 1061–1080.

Rayner, Keith, Gary E Raney, and Alexander Pollatsek (1995). "Eye movements and discourse processing." In: pp. 241–255.

Rayner, Keith, Arnold D Well, and Alexander Pollatsek (1980). "Asymmetry of the effective visual field in reading." In: *Perception & Psychophysics* 27.6, pp. 537–544.

Rayner, Keith, Sara C Sereno, Robin K Morris, A Rene Schmauder, and Charles Clifton Jr (1989). "Eye movements and on-line language comprehension processes." In: *Language and Cognitive Processes* 4.3-4, SI21–SI49.

Rei, Marek (2017). "Semi-supervised Multitask Learning for Sequence Labeling." In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vol. 1, pp. 2121–2130.

Rei, Marek and Anders Søgaard (2018). "Zero-shot Sequence Labeling: Transferring Knowledge from Sentences to Tokens." In: *Proceedings of the 16th Annual Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL 2018)*, pp. 293–302.

Rei, Marek and Helen Yannakoudakis (2016). "Compositional sequence labeling models for error detection in learner writing." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (ACL)*, pp. 1181–1191.

Rei, Marek and Helen Yannakoudakis (2017). "Auxiliary Objectives for Neural Error Detection Models." In: *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*, pp. 33–43.

Reichle, Erik D, Alexander Pollatsek, Donald L Fisher, and Keith Rayner (1998). "Toward a model of eye movement control in reading." In: *Psychological review* 105.1, p. 125.

Rohanian, Omid, Shiva Taslimipoor, Victoria Yaneva, and Le An Ha (2017). "Using Gaze Data to Predict Multiword Expressions." In:

Proceedings of the International Conference Recent Advances in Natural Language Processing (RANLP), pp. 601–609.

Rosenthal, Sara, Preslav Nakov, Svetlana Kiritchenko, Saif Mohammad, Alan Ritter, and Veselin Stoyanov (2015). "Semeval-2015 task 10: Sentiment analysis in Twitter." In: *Proceedings of the 9th international workshop on semantic evaluation (SemEval 2015)*, pp. 451–463.

Rotsztejn, Jonathan (2018). "Learning from Cognitive Features to Support Natural Language Processing Tasks." MA thesis. Switzerland: Eidgenössische Technische Hochschule Zürich.

Salojärvi, Jarkko, Ilpo Kojo, Jaana Simola, and Samuel Kaski (2003). "Can relevance be inferred from eye movements in information retrieval." In: *Proceedings of WSOM*. Vol. 3, pp. 261–266.

San Agustin, Javier, Henrik Skovsgaard, John Paulin Hansen, and Dan Witzner Hansen (2009). "Low-cost gaze interaction: ready to deliver the promises." In: *CHI'09 Extended Abstracts on Human Factors in Computing Systems*. ACM, pp. 4453–4458.

San Agustin, Javier, Henrik Skovsgaard, Emilie Mollenbach, Maria Barret, Martin Tall, Dan Witzner Hansen, and John Paulin Hansen (2010). "Evaluation of a low-cost open-source gaze tracker." In: *Proceedings of the 2010 Symposium on Eye-Tracking Research & Applications (ETRA)*. ACM, pp. 77–80.

Schmauder, A René, Robin K Morris, and David V Poynor (2000). "Lexical processing and text integration of function and content words: Evidence from priming and eye fixations." In: *Memory & Cognition* 28.7, pp. 1098–1108.

Schmidt, Anna and Michael Wiegand (2017). "A survey on hate speech detection using natural language processing." In: *Proceedings of the Fifth International Workshop on Natural Language Processing for Social Media*, pp. 1–10.

Sereno, Joan A and Allard Jongman (1997). "Processing of English inflectional morphology." In: *Memory & Cognition* 25.4, pp. 425–437.

Shain, Cory, Marten van Schijndel, Edward Gibson, and William Schuler (2016a). "Exploring memory and processing through a gold standard annotation of Dundee." In: *Proceedings of CUNY*.

Shain, Cory, Marten van Schijndel, Richard Futrell, Edward Gibson, and William Schuler (2016b). "Memory access during incremental sentence processing causes reading time latency." In: *Proceedings of the workshop on computational linguistics for linguistic complexity (CL4LC)*, pp. 49–58.

Shain, Cory, William Bryce, Lifeng Jin, Victoria Krakovna, Finale Doshi-Velez, Timothy Miller, William Schuler, and Lane Schwartz (2016c). "Modeling syntax acquisition via cognitively-constrained unsupervised grammar induction." In: *Proceedings of the 26th Interna-*

tional Conference on Computational Linguistics (COLING), pp. 964–975.

Shardlow, Matthew (2014). "Out in the Open: Finding and Categorising Errors in the Lexical Simplification Pipeline." In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 1583–1590.

Shigehalli, Vijayalaxmi and Vidya Shettar (2011). "Spectral Technique using Normalized Adjacency Matrices for Graph Matching." In: *International Journal of Computational Science and Mathematics* 3, pp. 371–378.

Singh, Abhinav Deep, Poojan Mehta, Samar Husain, and Rajkumar Rajakrishnan (2016). "Quantifying sentence complexity based on eye-tracking measures." In: *Proceedings of the workshop on computational linguistics for linguistic complexity (cl4lc)*, pp. 202–212.

Siyanova-Chanturia, Anna (2013). "Eye-tracking and ERPs in multi-word expression research: A state-of-the-art review of the method and findings." In: *The Mental Lexicon* 8.2, pp. 245–268.

Skovsgaard, Henrik, John Paulin Hansen, and Emilie Møllenbach (2013). "Gaze tracking through smartphones." In: *Gaze Interaction in the Post-WIMP World CHI 2013 One-day Workshop*.

Smith, Nathaniel J and Roger Levy (2010). "Fixation durations in first-pass reading reflect uncertainty about word identity." In: *CogSci*, pp. 1313–1318.

Smith, Samuel L, David HP Turban, Steven Hamblin, and Nils Y Hammerla (2017). "Offline bilingual word vectors, orthogonal transformations and the inverted softmax." In: *Proceedings of the 5th International Conference on Learning Representations (ICLR)*, pp. 1–9.

Socher, Richard, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D Manning, Andrew Ng, and Christopher Potts (2013). "Recursive deep models for semantic compositionality over a sentiment treebank." In: *Proceedings of the 2013 conference on empirical methods in natural language processing*, pp. 1631–1642.

Søgaard, Anders (2016). "Evaluating word embeddings with fMRI and eye-tracking." In: *Proceedings of the 1st Workshop on Evaluating Vector-Space Representations for NLP*, pp. 116–121.

Søgaard, Anders and Yoav Goldberg (2016). "Deep multi-task learning with low level tasks supervised at lower layers." In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*. Vol. 2, pp. 231–235.

Sparrow, Laurent, Sébastien Miellet, and Yann Coello (2003). "The effects of frequency and predictability on eye fixations in reading: An evaluation of the EZ Reader model." In: *Behavioral and Brain Sciences* 26.04, pp. 503–505.

Spitkovsky, Valentin Ilyich (2013). "Grammar Induction and Parsing with Dependency-and-Boundary Models." PhD thesis. Stanford University.

Staub, Adrian and Keith Rayner (2007). "Eye movements and on-line comprehension processes." In: *The Oxford Handbook of Psycholinguistics*. Ed. by Gareth Gaskell, pp. 327–342.

Stymne, Sara, Henrik Danielsson, Sofia Bremin, Hongzhan Hu, Johanna Karlsson, Anna Prytz Lillkull, and Martin Wester (2012). "Eye Tracking as a Tool for Machine Translation Error Analysis." In: *Proceedings of the International Conference on Language Resources and Evaluation (LREC)*, pp. 1121–1126.

Täckström, Oscar, Dipanjan Das, Slav Petrov, Ryan McDonald, and Joakim Nivre (2013). "Token and type constraints for cross-lingual part-of-speech tagging." In: *Transactions of the Association for Computational Linguistics (TACL)* 1, pp. 1–12.

Tamuz, Omer, Ce Liu, Serge Belongie, Ohad Shamir, and Adam Tauman Kalai (2011). "Adaptively learning the crowd kernel." In: *Proceedings of the 28th International Conference on Machine Learning (ICML)*, pp. 673–680.

Traxler, Matthew J (2002). "Plausibility and subcategorization preference in children's processing of temporarily ambiguous sentences: Evidence from self-paced reading." In: *The Quarterly Journal of Experimental Psychology: Section A* 55.1, pp. 75–96.

Traxler, Matthew, Robin Morris, and Rachel Seely (2002). "Processing subject and object relative clauses: Evidence from eye movements." In: *Journal of Memory and Language* 47, pp. 69–90.

Turian, Joseph, Lev Ratinov, and Yoshua Bengio (2010). "Word representations: a simple and general method for semi-supervised learning." In: *Proceedings of the 48th annual meeting of the Association for Computational Linguistics (ACL*, pp. 384–394.

Wallot, Sebastian, Beth A O'Brien, Guy Van Orden, C Westbury, G Jarema, and G Libben (2012). "Fractal and recurrence analysis of psycholinguistic data." In: *Methodological and Analytic Frontiers in Lexical Research. John Benjamins: Amsterdam*, pp. 395–430.

Wallot, Sebastian, Beth A O'brien, Anna Haussmann, Heidi Kloos, and Marlene S Lyby (2014). "The role of reading time complexity and reading speed in text comprehension." In: *Journal of Experimental Psychology: Learning, Memory, and Cognition* 40.6, pp. 1745–1765.

Wallot, Sebastian, Beth O'Brien, Charles A Coey, and Damian Kelty-Stephen (2015). "Power-law fluctuations in eye movements predict text comprehension during connected text reading." In: *Cognitive Science*, pp. 2583–2588.

Wang, Shaonan, Jiajun Zhang, and Chengqing Zong (2017). "Learning sentence representation with guidance of human attention."

In: *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence*, pp. 4137–4143.

Waseem, Zeerak (2016). "Are you a racist or am i seeing things? annotator influence on hate speech detection on twitter." In: *Proceedings of the first workshop on NLP and computational social science*, pp. 138–142.

Waseem, Zeerak and Dirk Hovy (2016). "Hateful symbols or hateful people? predictive features for hate speech detection on twitter." In: *Proceedings of the NAACL student research workshop*, pp. 88–93.

Weischedel, Ralph, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, et al. (2013). "Ontonotes release 5.0 LDC2013T19." In: *Linguistic Data Consortium*.

Wilson, Andrew, Christoph Dann, Chris Lucas, and Eric Xing (2015). "The human kernel." In: *Advances in neural information processing systems (NIPS)*, pp. 2854–2862.

Wilson, Theresa, Zornitsa Kozareva, Preslav Nakov, Sara Rosenthal, Veselin Stoyanov, and Alan Ritter (2013). "Sentiment analysis in Twitter." In: *Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pp. 312–320.

Wisniewski, Guillaume, Nicolas Pécheux, Souhir Gahbiche-Braham, and François Yvon (2014). "Cross-Lingual Part-of-Speech Tagging through Ambiguous Learning." In: *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Vol. 14, pp. 1779–1785.

Xu, P., K. A. Ehinger, Y. Zhang, A. Finkelstein, S. R. Kulkarni, and J. Xiao (2015). "TurkerGaze: Crowdsourcing Saliency with Webcam-based Eye Tracking." arXiv:1504.06755.

Yaneva, Victoria, Shiva Taslimipoor, Omid Rohanian, et al. (2017). "Cognitive Processing of Multiword Expressions in Native and Non-native Speakers of English: Evidence from Gaze Data." In: *International Conference on Computational and Corpus-Based Phraseology*. Springer, pp. 363–379.

Yannakoudakis, Helen, Ted Briscoe, and Ben Medlock (2011). "A new dataset and method for automatically grading ESOL texts." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL)*. Vol. 1. Association for Computational Linguistics, pp. 180–189.

Yimam, Seid Muhie, Chris Biemann, Shervin Malmasi, Gustavo Paetzold, Lucia Specia, Sanja Štajner, Anaïs Tack, and Marcos Zampieri (2018). "A Report on the Complex Word Identification Shared Task 2018." In: *Proceedings of the 13th Workshop on Innovative Use of NLP for Building Educational Applications (BEA)*. New Orleans, LA, USA: Association for Computational Linguistics.

Zelenina, Maria (2014). "Part of Speech Induction with Gaze Features." MA thesis. Edinburgh, UK: University of Edinburgh.

Zhang, Yue and Joakim Nivre (2011). "Transition-based dependency parsing with rich non-local features." In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics (ACL): Human Language Technologies*. Association for Computational Linguistics, pp. 188–193.

Zoph, Barret, Deniz Yuret, Jonathan May, and Kevin Knight (2016). "Transfer Learning for Low-Resource Neural Machine Translation." In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pp. 1568–1575.